Rijksuniversiteit Groningen **Faculteit der Wiskunde en Natuurwetenschappen**

Vakgroep Informatica

Formalizing Motivational Attitudes

Melle J. de Vries

begeleider: Prof.dr. G.R. Renardel de Lavalette

augustus 1996

RuG

Abstract

In this paper an introduction to the formalization of motivational attitudes of agents is given. The fundamentals for a logical characterization of motivational attitudes are reviewed. Further, an overview of the philosophical background concerning the explication of motivational attitudes (in relation to other mental attitudes) is provided. In the main part of this paper attention is paid to some formalisms in which some motivational attitudes are captured. Next, the scope is shifted from theory to practice. Some attempts to implement mentalistic notions are discussed. In order to let agents communicate with each other in a Multi-Agent System, some key concepts of Speech Act theory are combined with the formalization of motivational attitudes. In this paper, it is my aim to give an overview of some attempts in Artificial Intelligence to deal with motivational attitudes and to evaluate them critically.

Contents

Chapter 1

Introduction

1.1 Subject of this Paper

As can be deduced from the title the subject of this paper is the formalization of motivational attitudes. The formalization of motivational attitudes is part of the formalization of intelligent (or rational) agents which is a topic of continuing interest in *Artificial Intelligence*. Artificial Intelligence is the subfield of Computing Science which aims to construct agents that exhibit aspects of intelligent behavior. In this paper, I will focus my attention on the motivational attitudes of an intelligent agent.

1.1.1 What is an Agent?

According to [55], an agent is a computer system with the following properties:

1. An agent operates without the direct intervention of humans or others, and has some kind of control over its actions and internal state.
2. An agent can communicate to other agents via some kind of agent-communication-language.
3. An agent perceives his environment and responds in a timely fashion to changes that occur in it.
4. An agent does not simply act in response to his environment, he is able to exhibit goal-directed behavior by taking the initiative.
5. Agents are conceptualized or implemented using concepts that are more usually applied to humans.

For example, it is quite common in AI to characterize an agent using mentalistic notions like belief, knowledge, intentions, and even emotions [1].

Following McCarthy [35] and Dennett [15], a computer system is intelligent if you need to attribute cognitive concepts such as intentions and beliefs to it in order to characterize, understand, analyze, or predict its behavior. In order to design an intelligent agent the *intentional stance* is used very often. The intentional stance works as follows: "First you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have on the same considerations, and finally

you predict that this rational agent will act to further its goals in the light of its beliefs." [15, p.17]

Singh [48, p.3] provides us with the following reasons for adopting the intentional stance. Abstractions like intentions and belief are natural to humans who are not only the designers and analyzers of Multi-Agent Systems, but also the end users and requirements specifiers. Secondly, the abstractions provide succinct descriptions of, and help understand and explain, the behavior of complex systems. Further, they make available certain regularities and patterns of action that are independent of the exact physical implementation of the agents in the system. The abstractions may also be used by the agents themselves in reasoning about each other.

Among the abstractions one can distinguish between *informational* and *motivational* attitudes. Informational attitudes, like knowledge and belief, are related to the information that an agent has about the world it occupies. Motivational attitudes are those that in some way guide the agent's actions. They are sometimes called *pro-attitudes* to express the fact that these attitudes determine one's behavior in advance. You then have to think of desires, intentions, obligations, commitments, choices, etc. In this paper, however, I will mostly concentrate on the concept of intentions , because motivations reveal themselves very well in intentions.

"An *agent theory* must define how the attributes of agency are related. For example, it will need to show how an agent's information and pro-attitudes are related; how an agent's cognitive state changes over time; how the environment affects an agent's cognitive state; and how an agent's information and pro-attitudes lead it to perform actions. Giving a good account of these relationships is the most significant problem faced by agent theorists." [55, p.9]

1.1.2 What are Multi-Agent Systems?

Multi-Agent Systems are distributed computing systems that are designed as a collection of interacting autonomous agents, each having their own capacities and goals that are related to a common environment. The main goal of research concerning Multi-Agent Systems is to develop methods and instruments, like software architecture, reasoning models, and knowledge representation language, in order to enable an autonomous agent to coexist and cooperate with other agents. Singh [48, p.2] discerns two trends in Computing Science: The first one is the development of increasingly intelligent systems. Secondly, there is a trend towards the distribution of computing. The science of Multi-Agent Systems lies at the intersection of these trends. In this paper, I will mainly focus my attention on the single-agent case. Whenever possible, I will make an excursion to the multi-agent aspect. In Chapter 6 I will pay attention to the communication between intelligent agents to illustrate the application of motivational attitudes in rational interaction between agents.

1.2 Why Formalizing Motivational Attitudes?

During my study of Computing Science in Groningen, the emphasis has been laid on logics and specification. When my attention was attracted by Artificial Intelligence for the first time, it was soon suggested to me by Koen Hindriks, someone I studied with, to write a paper on the formalization of intelligent agents. I was interested in the logical background of Artificial Intelligence, and decided to restrict the subject to the formalization of motivational attitudes. Koen Hindriks concentrated mainly on the ability of agents. A result of his work can be found in [26]. During the third trimester in 1996 a seminar on logics for Multi-Agent

Systems was formed by Koen Hindriks, myself, and G.R. Renardel de Lavalette, our thesis supervisor.

When computer scientists want to write a program on something, they firstly have to formalize the various components of their objective. For Artificial Intelligence it is important to construct or analyze intelligent systems. In this paper, my main aim is to give an overview of some attempts to formalize motivational attitudes of intelligent agents in a multi-agent world. This may give us insight in the mental structure of people. Besides, it can convince us of the possibility of programming intelligent systems.

1.3 Outline of this Paper

In the following two chapters I will discuss the building blocks for a logic of motivational attitudes (especially intentions). In Chapter 2 some preliminary definitions concerning various logical systems are summed up. The elementary concepts of modal logic , epistemic logic, dynamic logic , and temporal logic are reviewed. In Chapter 3, the attention is directed at the theoretical basis underlying motivational attitudes. I will give a survey of the insights of philosophy and psychology concerning motivational attitudes. What do we mean, when we talk about someone's intentions? What is the meaning of an intention in opposition to other motivational attitudes? Which properties should be avoided in formalizing motivational attitudes? For the last question I will provide with formal definitions of some undesired properties. So, they can be used as criteria for the comparison of some formalizations of motivational attitudes.

In the fourth chapter I will compare some agent theories on intentions and other motivational attitudes. I take a look to the ideas of Cohen and Levesque [9], Rao and Georgeff [41], Konolige and Pollack [30], Singh [48], Huang et al.[27], and Van Linder et al.[33]. Firstly, I will discuss their formal frameworks. Then, I will look to some criticisms and use the criteria of Chapter 3 as a test. In the conclusions I will provide with three tables in which, respectively, the expressibility of the various formalisms, the temporal structure that is used in the theories, and the results of the test are summarized.

In the fifth chapter the emphasis is shifted from theory to practice. I will look how agents can be programmed using a specific language. An *agent language* is a system that allows one to program hardware or software computer systems in terms of some of the concepts discussed in the preceding chapters. I will describe AGENT-0, provided by Shoham, [46] and PLACA, provided by Thomas [50]. Further, I will discuss in this chapter if it is at all possible to program agents with the use of the concepts described in Chapter 4.

I come to look to the multi-agent world in the sixth chapter. In this chapter, I will describe how the individual attitudes can be extended to group attitudes. Single-agent knowledge is different from distributed knowledge [19]. Further, one can distinguish between intentions and collective intentions [23] [56]. I will also pay attention to the various *speech acts* [43] and the interaction between rational agents, because they convey interesting connections with motivational attitudes.

In the final chapter I will provide a summary of the results of my investigations.

Chapter 2

Fundamentals of a Logic for Motivational Attitudes

The main topic of this paper is the formalization of motivational attitudes of intelligent agents. In the introduction, I already said that I would concentrate on intentions as the main motivational attitude. In my opinion, intentions are distinct of other mental attitudes. Intentions are based on informational attitudes and provide a direction for achieving future states of the world. Intentions can move someone to act. Maybe, intentions can be characterized as follows: *Intentions are the necessary link between someone's knowledge and someone's actions in order to reach a future state of the world.*

In this chapter, I will provide some building blocks for a logic of intentions. For a good formalization of intentions several modal logics are needed. According to the above characterization of intentions there is a need for an epistemic logic [19], an action logic [24][31], and a temporal logic [18]. In the following sections I will describe some details of this logics. In the next chapter, I will discuss several properties of intentions.

2.1 Preliminary Definitions

In this section I will enumerate some basic topics from propositional, first order, and modal logics. For a thorough treatment of the various logics I refer the reader to [20]. Throughout this paper I will use the following sets as language constructs for the various logics.

- Φ is a nonempty set of primitive propositions, typically labeled $p, q, \ldots$
- $\mathcal{A}$ is a nonempty set of agent symbols, typically labeled $m, n, \ldots$
- $\mathcal{B}$ is a nonempty set of basic (or atomic) action symbols, typically labeled $a, b, \ldots$ Abstract or composed actions are denoted by $\alpha, \beta, \ldots$
- $\mathcal{X}$ is a set of arbitrary variables, typically labeled $x, y, \ldots$
- $\mathbf{T}$ is a set of (ordered) moments, typically labeled $t, t', \ldots$
- A set of connectives $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$.

2.1.1 Propositional Logic

Propositional logic deals only with the truth or falsity of propositional formulae. Propositional formulae are element of a language $\mathcal{L}$ and are composed of the set primitive propositions Φ with the help of the usual connectives. To denote propositional formulae I will use $\varphi, \psi, \ldots$ Further, the following abbreviations are used.

Convention 2.1.1.1 Abbreviations

1. $\mathsf{false} \stackrel{\mathsf{def}}{=} p \wedge \neg p$ for an arbitrary proposition p
2. $\mathsf{true} \stackrel{\mathsf{def}}{=} \neg\mathsf{false}$

2.1.2 First Order Logic

In first order logic, a domain of discussion D is assumed. It is possible to analyze propositions internally. Quantification over individual variables is permitted.

Definition 2.1.2.1 The language $\mathcal{L}$ is extended with formulae like $(\forall x \bullet \varphi)$ and $(\exists x \bullet \varphi)$. The formula φ can have the form $P(t_0, \ldots, t_n)$, where P is a predicate symbol and every t_i is an arbitrary variable, a constant or a function on variables. The n-ary predicate symbols $(P, Q, \ldots)$ are interpreted as concrete, n-ary relations over D, while the n-ary function symbols $(f, g, \ldots)$ are interpreted as concrete, n-ary functions on D.

2.1.3 Modal Logic

Modal logic includes modalities like *It is possible that ...* and *It is necessary that ...* In the sequel, $\Diamond$ is assumed to be the possibility operator, and $\Box$ is assumed to be the necessity operator.

Definition 2.1.3.1 Syntactic rules

Modal propositional logic contains all formulae of propositional logic and modal first order logic contains all formulae of first order logic. If φ is a formula of modal logic, then so are $\Diamond\varphi$ and $\Box\varphi$.

For the interpretation of modal formulae normally a possible-worlds semantics is used. Intuitively, this means that besides the true state of affairs, there are a number of other possible states of affairs (possible worlds). "In distributed computation application, a possible world may be seen as a possible global state of the system (that is, a possible combination of local states of the various processors) given a fixed protocol, and the worlds accessible to each agent from a given world consist of all global states in which its local state is the same as in the given world." [46, p.55]

Definition 2.1.3.2 Models for interpretation

$\mathcal{M} = \langle W, R, V \rangle$ is a model for interpretation of formulae of $\mathcal{L}$ if

1. W is set of possible worlds or states,
2. R is a binary relation on W, called an *accessibility relation*, and
3. V is an interpretation function.

Convention 2.1.3.3 A formula φ is said to be *valid*, denoted by $\models \varphi$, provided that for every structure $\mathcal{M}$ and every world w in $\mathcal{M}$ we have $\mathcal{M}, w \models \varphi$. A formula φ is said to be *valid in* $\mathcal{M}$, denoted by $\mathcal{M} \models \varphi$, provided that for every world w in $\mathcal{M}$ we have $\mathcal{M}, w \models \varphi$. A formula φ is said to be *satisfiable in* $\mathcal{M}$ iff there exists a world w with $\mathcal{M}, w \models \varphi$. A formula φ is said to be *satisfiable* there exists a structure $\mathcal{M}$ with φ satisfiable in $\mathcal{M}$.

Convention 2.1.3.4 In the sequel, the operator $\rightarrow$ is used to denote a logical implication. The operator $\Rightarrow$ is part of the metalanguage and is used to denote the application of a rule.

Definition 2.1.3.5 Semantic rules

1. $\mathcal{M}, w \models p$ iff $V_{\mathcal{M}}(p) = 1$
2. $\mathcal{M}, w \models P(t_1, \ldots, t_n)$ iff $\langle V_{\mathcal{M}}(t_1), \ldots, V_{\mathcal{M}}(t_n) \rangle \in V_{\mathcal{M}}(P)$
3. $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$
4. $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
5. $\mathcal{M}, w \models \varphi \vee \psi$ iff $\mathcal{M}, w \models \varphi$ or $\mathcal{M}, w \models \psi$
6. $\mathcal{M}, w \models \varphi \rightarrow \psi$ iff $\mathcal{M}, w \models \neg\varphi$ or $\mathcal{M}, w \models \psi$
7. $\mathcal{M}, w \models (\forall x \bullet \varphi(x))$ iff $\mathcal{M}, w \models \varphi(d)$, for all $d \in D$
8. $\mathcal{M}, w \models (\exists x \bullet \varphi(x))$ iff $\mathcal{M}, w \models \varphi(d)$, for some $d \in D$
9. $\mathcal{M}, w \models \Box\varphi$ iff $(\forall w' \in W \bullet (w, w') \in R \Rightarrow \mathcal{M}, w' \models \varphi)$
10. $\mathcal{M}, w \models \Diamond\varphi$ iff $(\exists w' \in W \bullet (w, w') \in R \Rightarrow \mathcal{M}, w' \models \varphi)$

There are several extensions to modal logic with almost the same set of semantic rules. The difference is that the extensions depart from adapted accessibility relations. In *epistemic logic* an epistemic accessibility relation is used. In *dynamic logic* programs or abstract actions are interpreted as binary relations on states. *Temporal logic* deals with the accessibility of future states of the world.

Definition 2.1.3.6 Binary relations

1. A relation R is *reflexive* iff $(\forall x \bullet (x, x) \in R)$
2. A relation R is *transitive* iff $(\forall x, y, z \bullet (x, y) \in R \wedge (y, z) \in R \Rightarrow (x, z) \in R)$
3. A relation R is *symmetric* iff $(\forall x, y \bullet (x, y) \in R \Rightarrow (y, x) \in R)$
4. A relation R is *serial* iff $(\forall x \bullet (\exists y \bullet (x, y) \in R))$
5. A relation R is *euclidean* iff $(\forall x, y, z \bullet (x, y) \in R \wedge (x, z) \in R \Rightarrow (y, z) \in R)$
6. A relation R is called an *equivalence relation* iff R is reflexive, transitive, and symmetric.

Every extension of normal modal logic can be characterized by one or more axioms. Every axiom corresponds to one of the accessibility-relations. There are two axioms that hold for every modal logic, namely the K-axiom (distributivity) and the N-axiom (necessitation rule).

Definition 2.1.3.7 Modal axioms

K: $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$

N: $\varphi \Rightarrow \Box\varphi$

T: $\Box\varphi \rightarrow \varphi$ corresponds to the reflexive relations

4: $\Box\varphi \rightarrow \Box\Box\varphi$ corresponds to the transitive relations

B: $\varphi \rightarrow \Box\Diamond\varphi$ corresponds to the symmetric relations

D: $\Box\varphi \rightarrow \Diamond\varphi$ corresponds to the serial relations

5: $\Diamond\varphi \rightarrow \Box\Diamond\varphi$ corresponds to the euclidean relations

Often, some axioms are combined in order to obtain the desired logic. For example, $S5$ corresponds with the class of all equivalence relations, i.e., all reflexive, transitive, and symmetric relations. $S5$ is thus a combination of the T, 4, and 5 axioms.

2.2 Reasoning about Knowledge and Belief

2.2.1 Epistemic Logic

Epistemic logic is the logic of the concepts of knowledge. Epistemic logic is just a variant of modal logic. Normally, the framework for modeling knowledge is based on possible worlds. An agent is said to know a fact φ, denoted by $K(\varphi)$, if φ is true at all the worlds he considers possible. Models for epistemic logic resemble the models of ordinary modal logic. Instead of the accessibility relation R of modal logic, I define $\mathcal{E}$ to be the *epistemic* accessibility relation. $(w, w') \in \mathcal{E}$ means that the agent considers world w' possible, given his information in world w. It seems natural to make $\mathcal{E}$ an equivalence relation. So, $\mathcal{E}$ is reflexive, symmetric, and transitive.

Definition 2.2.1.1 Semantic rule for the knowledge operator
$\mathcal{M}, w \models K(\varphi)$ iff $(\forall w' \bullet (w, w') \in \mathcal{E} \Rightarrow \mathcal{M}, w' \models \varphi)$

The following property, occasionally called the *Knowledge Axiom*, has been taken by philosophers to be the major one distinguishing knowledge from belief [19, p.32]. The Knowledge Axiom corresponds to the T-axiom defined in the previous section. If $K(\varphi)$ holds at a particular world $\mathcal{M}, w$ and the epistemic accessibility relation $\mathcal{E}$ is reflexive, then φ is true at all worlds that the agent considers possible, so in particular it is true at $\mathcal{M}, w$. Usually, the accessibility relation for beliefs is taken to be not reflexive.

Lemma 2.2.1.2 $\models K(\varphi) \rightarrow \varphi$

As described in chapter 4, for the formalization of motivational attitudes of rational agents most computer scientists use the concept of belief instead of knowledge. As modal operator for belief, BEL is used. For a belief system, the accessibility relations are serial, transitive, and euclidean. A logic of belief is sometimes called a *doxastic logic* instead of an epistemic logic.

Some other properties that hold for the above definition of knowledge follow from the possible world approach that is chosen. Those properties are stated in the next lemma.

Lemma 2.2.1.3 For all formulae φ, ψ and all structures $\mathcal{M}$ where each accessibility relation is an equivalence relation

1. $\mathcal{M} \models (K(\varphi) \wedge K(\varphi \rightarrow \psi)) \rightarrow K(\psi)$
2. $\mathcal{M} \models \varphi \Rightarrow \mathcal{M} \models K(\varphi)$
3. $\mathcal{M} \models K(\varphi) \rightarrow KK(\varphi)$
4. $\mathcal{M} \models \neg K(\varphi) \rightarrow K\neg K(\varphi)$

The first property states that the knowledge operator distributes over implication. The second property formalizes the fact that if φ is true at all the possible worlds of structure $\mathcal{M}$, then φ must be true at all the worlds that an agent considers possible in any given world in $\mathcal{M}$. The last two properties say that agents can do introspection regarding their knowledge. They know what they know and what they do not know. The above properties correspond respectively to the K, N, 4, and 5 axioms of normal modal logic. So, the epistemic accessibility relation $\mathcal{E}$ is transitive, euclidean, and reflexive (lemma 2.2.1.2). This is just another way to say that $\mathcal{E}$ is an equivalence relation. In the literature this logic is referred to as $S5$.

2.2.2 Logical Omniscience

One of the main drawbacks of modeling knowledge using a variant of modal logic, is that the agents are assumed to be *logically omniscient*, i.e., the agents know all the consequences of their knowledge and they know all tautologies [19, p.309]. However, if we consider human reasoning, then we have to say that people are simply not logically omniscient. For example, a person can know the rules of chess without knowing whether or not White has a winning strategy.

Definition 2.2.2.1 The term logical omniscience actually refers to a family of related closure conditions.

1. *Knowledge of valid formulae*: If φ is valid, then every agent knows φ.
2. *Closure under logical implication*: If an agent knows φ and if φ logically implies ψ, then he knows ψ.
3. *Closure under logical equivalence*: If an agent knows φ and if φ and ψ are logically equivalent, then he knows ψ.

In [19, pp.313-346] a number of different approaches to avoiding or alleviating the logical omniscience problem is suggested. For example, the *awareness approach* adds awareness as another component of knowledge, contending that one cannot explicitly know a fact unless one is aware of it. Systems of explicit knowledge (or belief) are more suited to modeling finite agents [14]. *Impossible worlds* are sometimes introduced to allow inconsistent formulae to be true. Moreover, not all valid formulae need be true.

Almost all attempts in the literature to solve the problem of logical omniscience consist in weakening the standard epistemic systems. However, according to Duc [17], this solution is not satisfactory. In this way logical omniscience can be avoided, but many intuitions about the concepts of knowledge and belief get lost. To prevent this Duc proposes to 'temporalize' epistemic logic. He introduces dynamic logic in epistemic logic to express the fact that one needs time to perform an inference. $[R_i]K(\varphi)$ has the following meaning: always after using rule R_i the agent knows φ. $\langle R_i \rangle K(\varphi)$ formalizes the fact that sometimes after using rule R_i the agent knows φ.

2.3 Reasoning about Actions

2.3.1 Propositional Dynamic Logic

Propositional Dynamic Logic describes the properties of the interaction between programs and propositions that are independent of the domain of computation. A program can be viewed as a transformation of states. Given an initial (input) state, the program will go through series of intermediate states, perhaps eventually halting in a final (output) state. A sequence of states that can be obtained from the execution of a particular program starting from a given input state is called a *trace*. Traces can be finite or infinite. They need not be uniquely determined by their start state, because nondeterministic programs are allowed.

Definition 2.3.1.1 Syntactic rules

Compound propositions and programs are defined by mutual induction, as follows. If φ, ψ are propositions and α, β are programs, then

1. $(\neg\varphi), (\varphi \vee \psi), (\varphi \wedge \psi), (\varphi \rightarrow \psi), (\varphi \leftrightarrow \psi)$, and $(\langle\alpha\rangle\varphi)$ are propositions
2. $(\alpha;\beta), (\alpha + \beta), (\alpha^*)$, and $(\varphi?)$ are programs

Parentheses can be omitted. The intuitive meaning of $\langle\alpha\rangle\varphi$ is that it is possible to execute α and terminate in a state satisfying φ. Further, $(\alpha;\beta)$ is the sequential composition of α and β, $(\alpha + \beta)$ is the nondeterministic choice between α and β, (α^*) is the reflexive and transitive closure of state transition relations realized by α, and $(\varphi?)$ is a test condition on φ. $[\alpha]\varphi$ is an abbreviation of $(\neg\langle\alpha\rangle\neg\varphi)$ and has the following intuitive meaning: whenever α terminates, it must do so in a state satisfying φ. For fixed program α, the operator $[\alpha]\varphi$ behaves like a modal necessity operator and the operator $\langle\alpha\rangle\varphi$ behaves like a possibility operator of modal logic.

Definition 2.3.1.2 Models for interpretation

1. A model $\mathcal{M} = \langle W, I\rangle$ consists of an abstract set of states W and an interpretation function I.
2. Each proposition φ is interpreted as a subset of W
3. Each program α is interpreted as a binary relation on W

If p is an atomic proposition, then $I(p) \subseteq W$. If a is an atomic program symbol, then $I(a) \subseteq W \times W$. I' is the extension of I, such that I^* is also defined on compound programs and propositions.

Definition 2.3.1.3 Interpretation of formulae and programs

1. $I'(\langle\alpha\rangle\varphi) = \{w \in W | (\exists w' \in W \bullet (w, w') \in I'(\alpha) \wedge w' \in I'(\varphi))\}$
2. $I'(\alpha;\beta) = I'(\alpha) \circ I'(\beta) = \{(w, w'') | (\exists w' \in W \bullet (w, w') \in I'(\alpha) \wedge (w', w'') \in I'(\beta))\}$
3. $I'(\alpha + \beta) = I'(\alpha) \cup I'(\beta) = \{(w, w') | (w, w') \in I'(\alpha)$ or $(w, w') \in I'(\beta)\}$
4. $I'(\alpha^*) = \bigcup_{n \geq 0} (I'(\alpha))^n$, where $(I'(\alpha))^0 = \{(w, w) | w \in W\}$ and $(I'(\alpha))^{n+1} = (I'(\alpha)) \circ (I'(\alpha))^n$
5. $I'(\varphi?) = \{(w, w) | w \in I'(\varphi)\}$

Definition 2.3.1.4 Abbreviations

1. $[\alpha]\varphi \stackrel{\text{def}}{=} \neg\langle\alpha\rangle\neg\varphi$
2. **skip** $\stackrel{\text{def}}{=}$ (true?)
3. **fail** $\stackrel{\text{def}}{=}$ (false?)
4. **if** φ **then** α **else** $\beta \stackrel{\text{def}}{=} (\varphi?;\alpha + \neg\varphi?;\beta)$
5. **while** φ **do** $\alpha \stackrel{\text{def}}{=} ((\varphi?;\alpha)^*;\neg\varphi?)$

2.3.2 First Order Dynamic Logic

The main difference between First Order Dynamic Logic and the Propositional variant is the presence of a first order structure D, called the *domain of computation*, over which first order quantification is allowed. States are no longer abstract points, but *valuations* of a set of variables over D. Primitive programs are no longer abstract binary relations, but *assignments* of the form $x := t$, for example where x is a variable and t is a term. Primitive assertions are now first order formulae. I refer the interested reader to [24].

2.3.3 Ability

Sometimes an *ability*-operator, A, is added to the action theory [33]. The formula $\mathrm{A}(\alpha)$ denotes the fact that the agent has the ability to do α. If an agent has at his disposal only beliefs and intentions concerning a particular condition, but lacks the ability to bring about that condition, he will never reach that condition. Singh introduces the term *know-how*: "An agent knows how to achieve φ, if he is able to bring about the conditions for φ through his actions." [48, p.85] Because this paper deals with motivational attitudes, I will not focus on this topic.

2.4 Reasoning about Time

Temporal Logic provides a formal system for qualitatively describing and reasoning about how the truth values of assertions change over time. There are various alternative systems of temporal logic. One can distinguish between *propositional* and *first order* temporal logic. Another distinction concerns the view regarding the underlying nature of time. Time can be modeled as a *linear* time-line and as a *branching-time* structure. In the last case, time may split into alternate courses representing different possible futures. Systems for temporal logic may be *endogeneous*, in which case all temporal operators are interpreted in a single universe corresponding to a single concurrent program, or *exogeneous* to allow expression of correctness properties concerning several different programs in the same formula. Temporal operators can be evaluated as true or false of *points* in time or over *intervals* of time. Time structures may be *discrete* or *continuous*. For the first variant the nonnegative integers serve as temporal structure, for the latter reals (or rationals) are used. In most temporal logics only *future*-tense operators are provided. However, sometimes *past*-tense operators are added. In the following subsections I will discuss the linear variant and the branching variant of temporal logic. I will discuss only the propositional version. To obtain first order temporal logic just take propositional temporal logic and add to it a first order language.

2.4.1 Linear Temporal Logic

In the following formalization, it is assumed that time is discrete , has an initial moment with no predecessors, and is finite into the future. The symbol $\sigma = (w_0, w_1, w_2, \ldots) = (\sigma(0), \sigma(1), \sigma(2), \ldots)$ is used to denote a timeline as an infinite sequence of states. Propositional logic is extended to propositional linear temporal logic by introducing the following operators.

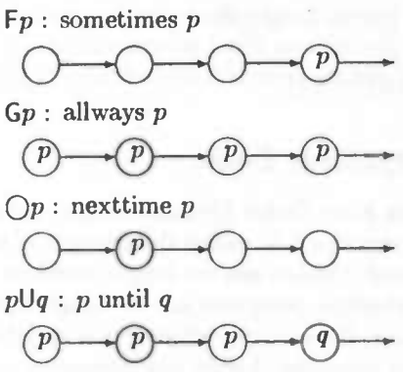

Figure 2.1: Intuition for linear-time operators

Definition 2.4.1.1 Syntactic rules
If φ and ψ are formulae of $\mathcal{L}$, then

1. $\varphi \mathsf{U} \psi$ (φ until ψ),
2. $\bigcirc \varphi$ (nexttime φ),
3. $\mathsf{F}\varphi \equiv \mathsf{true}\mathsf{U}\varphi$ (sometime φ), and
4. $\mathsf{G}\varphi \equiv \neg\mathsf{F}\neg\varphi$ (always φ) are formulae of $\mathcal{L}$.

The modality $\varphi \mathsf{U} \psi$ asserts that ψ does eventually hold and that φ will hold everywhere prior to ψ. The modality $\bigcirc\varphi$ holds now iff φ holds at the next moment. $\mathsf{F}\varphi$ means that at some future moment φ is true. $\mathsf{G}\varphi$ means that at all future moments φ is true. For the formal semantic rules a notational convention is used: σ^i denotes the suffix path $(w_i, w_{i+1}, w_{i+2}, \ldots)$.

Definition 2.4.1.2 Semantic rules

1. $\mathcal{M}, \sigma \models \varphi \mathsf{U} \psi$ iff $(\exists j \bullet \mathcal{M}, \sigma^j \models \psi$ and $(\forall i \bullet 0 \leq i < j \Rightarrow \mathcal{M}, \sigma^i \models \varphi))$
2. $\mathcal{M}, \sigma \models \bigcirc\varphi$ iff $\mathcal{M}, \sigma^1 \models \varphi$
3. $\mathcal{M}, \sigma \models \mathsf{F}\varphi$ iff $(\exists j \bullet \mathcal{M}, \sigma^j \models \varphi)$
4. $\mathcal{M}, \sigma \models \mathsf{G}\varphi$ iff $(\forall j \bullet \mathcal{M}, \sigma^j \models \varphi)$

Several interesting validities can be deduced from the above definitions, but I omit them here and refer the reader to [18].

2.4.2 Branching Temporal Logic

In branching-time temporal logics the structure of time corresponds to an infinite tree. It is allowed that a node in the tree has infinitely many successors, while it is required for each node to have at least one successor. In the following, state formulae are formulae that can be true or false of states (moments, worlds). Path formulae can be true or false of paths. CTL^* is the full branching-time logic, which consists of a set of state formulae generated by the next rules.

Definition 2.4.2.1 Syntactic rules

1. Each atomic proposition p is a state formula.
2. If φ and ψ are state formulae, then so are $\neg\varphi$, $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \rightarrow \psi$.
3. If φ is a path formula, then $\mathsf{E}\varphi$ and $\mathsf{A}\varphi$ are state formulae.
4. Each state formula is also a path formula.
5. If φ and ψ are path formulae, then so are $\neg\varphi$, $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \rightarrow \psi$.
6. If φ and ψ are path formulae, then so are $\varphi \mathsf{U} \psi$, $\bigcirc\varphi$, $\mathsf{F}\varphi$, and $\mathsf{G}\varphi$.

Definition 2.4.2.2 Models for interpretation
$\mathcal{M} = \langle W, \prec, V \rangle$ is a modal for interpretation of formulae of $\mathcal{L}$ if

1. W is a set of states (worlds),
2. $\prec$ is a total binary relation on W,
3. V is an interpretation function,
4. $\prec$ contains no directed cycles,
5. each $w \in W$ has at most one $\prec$-predecessor (no "merging of paths"), and
6. there exists a unique $w \in W$ -called the root- from which all other states (worlds) in W are reachable and that has no $\prec$-predecessors.

$\mathcal{M}, w_0 \models \varphi$ means that state formula φ is true in structure $\mathcal{M}$ at state w_0. $\mathcal{M}, \sigma \models \varphi$ means that path formula φ is true in structure $\mathcal{M}$ of fullpath σ. There are two new time operators. The quantifier $\mathsf{A}\varphi$ means that for all paths (or branches) in the future φ will hold (*inevitable*(φ)). Analogously, $\mathsf{E}\varphi$ means that sometime in the future φ will hold (*optional*(φ)).

Definition 2.4.2.3 Semantic rules

1. $\mathcal{M}, w_0 \models \mathsf{E}\varphi$ iff $(\exists \sigma = (w_0, w_1, w_2, \ldots) \in \mathcal{M} \bullet \mathcal{M}, \sigma \models \varphi)$
2. $\mathcal{M}, w_0 \models \mathsf{A}\varphi$ iff $(\forall \sigma = (w_0, \ldots) \in \mathcal{M} \bullet \mathcal{M}, \sigma \models \varphi)$
3. $\mathcal{M}, \sigma \models \varphi$ iff $\mathcal{M}, w_0 \models \varphi$, where σ begins at w_0.

2.5 Conclusions

In this chapter the building blocks for a logic of motivational attitudes (especially intentions) are provided. In the beginning of this chapter, I defined intentions to be the necessary link between someone's knowledge and someone's actions in order to reach a future state of the world. Thus, a logic of knowledge (or belief) and a logic of action are needed. Therefore, I formulated some key concepts of epistemic logic and dynamic logic. In order to deal with changing motivational attitudes, I added some key concepts of temporal logic. Most theories of rational agents are builded on the basic definitions provided in this chapter. Sometimes, another notation is used. As far as possible I will use throughout this paper the notation as introduced above. Therefore, I have to adapt sometimes the original notation of the various theories.

Chapter 3

Properties of Motivational Attitudes

In order to formalize motivational attitudes in such a way that it links up with our pretheoretic understanding it is helpful to list some logical and intuitive properties that motivational attitudes may, or may not, be taken to have. In my opinion, intentions form the most important motivational attitude. Thus, I will mainly focus my attention on intentions. Singh [48, pp.55-63] discusses a list of thirteen dimensions of variation in the study of intentions. I will describe them to provide the reader a philosophical background concerning intentions.

In this chapter, I will also discuss other motivational attitudes, such as goals, preferences, desires, wishes, choices, commitments, etc. I will investigate how these attitudes are related to each other and how they can be distinguished from each other.

In the literature on motivational attitudes like intentions, some authors have tackled the desirability of some possible counterintuitive properties that can be deduced from formalizations of intentions [5]. Computer scientists have tried to meet the desiderata of a formalization of intentions. But sometimes it is hard to connect theory and practice. At the end of this chapter, I will give some criteria for a good formalization of intentions.

3.1 Philosophical Background

[Dim-1] Intentions can variously be taken to be towards (a) *propositions* that an agent is deemed to intend to achieve or (b) *actions* that an agent is deemed to intend to perform. In the approach of Singh intentions per se are taken to apply to propositions, which makes for a natural discussion of their logical properties. Cohen and Levesque [9] have formalized both variants of intention.

[Dim-2] Intentionality is directedness [44]. We can distinguish between *future-directed* and *present-directed* intentions. In the first case intentions are taken to be towards future states of the world or future actions and in the second case intentions are taken to be towards present actions. A related distinction is that between *achievement-goals* and *maintenance-goals* [27]. I think, it is preferable to use the term intention in the first sense.

[Dim-3] There is also a difference between *intending something* and *doing it intentionally*. If an agent intends to do a particular action, then he must have a prior intention to

do that action. If an agent is doing something intentionally, however, he is purposefully performing it, but not with any prior intention to do so. In order to explain one's behavior we need the notion of prior intention, because only the conditions that are intended may be used to explain an agent's actions. So, I will restrict intended conditions to the former sense. In the other case the conditions occur as, possibly contingent, consequences of following a strategy.

[Dim-4] It is assumed that an agent believes that his intentions are *satisfiable*. This assumption is based on another assumption: agents are rational in some sense.

[Dim-5] We can also assume that an agent's intentions are *mutually consistent*. Inconsistency among an agent's intentions would make his mental state too incoherent for him to act.

[Dim-6] For the purposes of designing and analyzing intelligent systems, it may be acceptable to let one's theory validate the *closure under logical consequence* although in general intentions are not closed under logical consequence. So, mathematical reasoning contrasts here with our intuitions about a complicated concept such as intentions.

[Dim-7] We can also say that in general intentions are not *closed* even *under beliefs*. For example, an agent may intend φ, believe that φ necessarily entails ψ and not intend ψ. The most cited example is that of the the agent going to the dentist. The agent has the intention to have a tooth filled without also having the intention to suffer pain, although the agent believes that it is inevitable that pain always accompanies having a tooth filled. The inference aimed at is also termed the *side-effect problem* [40].

[Dim-8] On the other side, if an agent a intends φ, and believes that ψ is a necessary means to φ, a should intend ψ. So intentions are *closed under means* [3, p.126]. If I have the intention to have a tooth filled, I should also have the intention to go to the dentist.

[Dim-9] A property of intentions is that they usually involve some measure of *commitment* on part of the agent. That is, an agent who has an intention is committed to achieving it and will persist with it through changing circumstances. But that can not hold unconditionally. Cohen and Levesque [9] define an intention to be a persistent goal. In their formalism, an agent has a persistent goal if he will not give up the goal before he believes that the goal is realized or will never be realizable. Van Linder et al. [33] define a special action for an agent committing himself to an action. Rao and Georgeff [22] have formalized the process of *intention maintenance* in the context of changing beliefs and desires. Intended situations have to be dropped if they are not consistent anymore with beliefs or desires.

[Dim-10] We can say that intentions are *causes of actions* by agents. From this point of view we can conceptually differentiate intentions from desires and beliefs.

[Dim-11] It should be inconsistent for an agent to intend a proposition φ and simultaneously believe that φ will not occur. So, it makes sense to assume that an agent's *intentions are consistent with* his *beliefs* about the future.

[Dim-12] *Intentions do*, however, *not entail beliefs*. It should be consistent for an agent to intend φ and yet not believe that φ will occur. An agent may intend to reach the top of Mount Everest, but does not necessarily have to believe that he will succeed. The principles in Dim-11 en Dim-12 are also called the *intention-belief-inconsistency* and the *intention-belief-incompleteness* These two principles put together is called the

asymmetry-thesis [40]. Pears introduces the concept of probability: "A minimal future factual belief is an essential part of every intention. The agent must believe that his intention to perform a φ action makes it probable that he will perform one. The probability may be very low, but he must believe that it exists and that his intention confers it on his performance." [36, pp.78f]

[Dim-13] *Intentions are* usually taken to be *distinct from beliefs*, although they are always taken to be related to them. In most formalisms concerning motivational attitudes, beliefs and intentions are independently defined and related to each other by a sort of *realism constraint*, meaning that each intention has to be supported by the beliefs of the agent.

So far the discussion of some dimensions of variation concerning intentions. In the following section, I will discuss how intentions are related to other motivational attitudes.

3.2 Intentions in Opposition to other Motivational Attitudes

In this section I will provide a discussion of some motivational attitudes and their difference with intentions. I will discuss successively willing, preferences, desires, wishes, goals, commitments, and choices. Davidson [13, p.102] argues that intentions form a subclass of the motivational attitudes:

> Wants, desires, principles, prejudices, felt duties, and obligations provide reasons for actions and intentions, and are expressed by prima facie judgements; intentions and the judgements that go with intentional actions are distinguished by their all-out or unconditional form. Pure intendings constitute a subclass of the all-out judgements, those directed to future actions of the agent, and made in the light of beliefs.

3.2.1 Willing

Harman [25] distinguishes two notions of intentionality. He uses the term *intending* and related terms for the stronger notion, so that intending in this sense does involve believing one will do as one intends. He uses the term *willing* for the weaker notion, so that willing does not involve believing one will do as one wills. The notion of willing is, in this point of view, conceptually more basic than the notion of intending, but willing of the sort that can initiate action can occur only as a component of intending and only if one believes one will do as one wills.

Haddadi [23] uses the concepts of *willing* and *want* as decisions that form the necessary links between an agent's goals and the resulting intentions. $(Willing\ m\ \varphi)$ means that agent m is willing to achieve the condition φ individually, i.e., he has chosen a plan to achieve φ. $(Want\ m\ n\ \varphi)$ means that agent m has chosen agent n to achieve φ. The concepts of *willing* and *want* resemble in some sense the notion of *choice* that will be discussed further.

3.2.2 Preferences

Preferring something is comparative. You can prefer a proposition or an action to another. In [27] [28], and [33] preferences form the basis for the formalization of rational (or intelligent) agents. Preferences of an agent describe aspects of states of affairs that the agent prefers

to be the case. Van Linder et al. [33] distinguish between *implicit* and *explicit* preferences. Each agent has a possibly partial preference order at his disposal. The implicit preferences are those situations that the agent considers preferable to the current situation. Explicit preferences are distinct preferences that the agent made knowable. According to Van Linder et al. an agent's preferences consist of both an implicit and an explicit component. The semantics for preferences that they present is as follows: an agent prefers some condition iff it is true at a set of preferred alternatives, i.e. the agent implicitly prefers the formula, and it is furthermore considered to be an explicit preference.

Huang et al. [27] also consider preferences as the basis for rational action. In the language of their formal framework they define a binary preference-operator. Preferences are relations between propositions. Every agent has a preference order. However, both formalizations of preferences lack the connection with actions. The reason may be obvious. When some agent prefers something (to something else), he does not necessarily want to realize it. When some agent intends something, it may be expected that the agent wants to realize it (Dim-10). A discussion of preference logics can be found in [28, pp.99-117].

3.2.3 Desires

Brand [3, pp.123-127] provides us five differences between intending and desiring.

- The strength of a desire can change over time, but not so for an intention.
- Related to the first difference, desiring can be scaled in strength, but not intending.
- It is possible for a normal person to have incompatible desires but it is not possible for him to have incompatible intentions.
- It is far from extraordinary when someone desires to achieve some end but does not desire to do what is necessary to achieve it. Intentions are closed under necessary means, as stated in Dim-8.
- Intending is more closely connected with action than desiring. A person can desire to do something himself or he can desire that someone else do something. But a person can only intend to do something himself.

In my opinion intentions can also be scaled in strength and their strength can also change over time [16]. So, I do not fully agree with Brand, but I return to this topic later on in this paper.

Bratman [6, p.22] distinguishes between two kinds of *pro-attitudes*. Pro-attitudes play a motivational role: in concert with belief they can move us to act. While desires are said to be *potential influencers of conduct*, intentions are said to be *conduct-controlling* pro-attitudes.

3.2.4 Wishes

In [34, p.150f] Van Linder defines wishes to be the primitive motivational attitude that models the things that an agent likes to be the case. Wishes range over propositions, which corresponds to the idea that agents wish for certain aspects of the world. In my opinion, wishes behave like desires and do not constitute a separate motivational attitude, since they can be inconsistent and incompatible given the agent's resources.

3.2.5 Goals

Goals are often considered as primitives that determine what an agent seeks to achieve. In [33] it is expressed thus: goals are unfulfilled, realistic preferences. Cohen and Levesque define the set of goals as a set of consistent desires [10]. Rao and Georgeff define goals to be chosen desires of the agent that are consistent and achievable [41]. Huang et al. discuss four goal definitions and some obvious modifications [27]. A *good goal* is a situation that is preferred to its negation. A *satisficing goal* is a preferred situation that is made accessible via a search. A *maximal goal* is an accessible situation to which no other accessible situation is preferred. And finally, an *optimal goal* is a unique maximal goal. Further, they distinguish between *achievement* goals and *maintenance* goals. The former denote that a situation is not reached yet. The latter denote that the preferred situation is reached and has to be maintained.

Goals approximate the notion of intentions. Some authors use goals as the main motivational attitude. The most authors, however, define intentions as a separate concept. The most important reason to distinguish goals from intentions is the fact that the latter involve some measure of commitment (Dim-9). You will not easily give up your intentions.

3.2.6 Commitments

As mentioned above, commitments are part of intentions. An agent who has an intention is committed to achieving it and will persist with it through changing circumstances. However, an agent will not remain committed to an intention in all possible circumstances. There will be occasions when it is obviously irrational for the agent to remain committed, for example, the agent may come to believe the intention is impossible to achieve.

In [5] three dispositions of the commitment aspect of intention are discussed. 1) An agent will tend to retain an intention without reconsideration – because agents are resource bounded they can not constantly reconsider the merits of their intention. 2) An agent tends to reason from the intention to (sub-)intentions which play a part in the agent's plan. For example, an agent will reason from an intention to sub-intentions concerning mere specific actions. 3) An agent tends to reason in a way which constrains the adoption of sub-intentions, so that possible courses of action incompatible with the intention are not seriously considered.

Dongha [16] argues that there is a tension between dropping an intention and remaining committed to one when situated in an unpredictable environment. To overcome this tension, he designs a commitment mechanism for intentions. With each intention a *commitment level* is associated as a measure of anticipated and invested resources. On this basis a priority ordering for intentions is defined to reconsider them in a changing environment in that order.

Cohen and Levesque [9] use a definition of intention with a built-in definition of commitment. As Cohen and Levesque point out, their definition of persistent goal leads to fanaticism. Rao and Georgeff [41] introduce three commitment strategies. A *blind agent* maintains his intentions until he believes that he has actually achieved them. A *single-minded agent* maintains his intentions as long as he believes that they are still realizable. An *open minded agent* maintains his intentions as long as they are still among his goals. In [42, p.320] Rao and Georgeff argue as follows: "A commitment usually has two parts to it: one is the condition that the agent is committed to maintain, called the *commitment condition*, and the second is the condition under which the agent gives up the commitment, called the *termination condition*. More formally, we define a commitment operator C as follows: $\varphi_1 C \varphi_2 \equiv \mathsf{A}(\varphi_1 \mathsf{U} \varphi_2)$ where φ_1 is a commitment condition and φ_2 is a termination condition."

In an article on changing attitudes, Bell [2] introduces persistence rules for beliefs, desires, and intentions. Agents persist with their attitudes unless they have reason to change them. Agents are not continually reasoning about and revising their intentions. Reconsideration should be the exception, not the rule.

A last remark on commitments. I take them to be the result of inner motivation in opposition to one's obligations. The latter are in my opinion the result of outer motivation. It can be obligatory for a person to perform some action. However, Shoham [46] is the only author that paid attention to this fact of obligation.

3.2.7 Choices

Shoham [46] defines a choice (or decision) as a commitment to oneself. Intentions are, however, not part of his work. So, may be intentions can be identified with the concept of choice in the article of Shoham. Choices are, in my opinion, the underlying concept of intentions. An agent will not intend a particular condition, if he has not the possibility to choose for another condition. Bratman distinguishes between choice and intention by saying [6, p.29]:

> We should distinguish what is chosen on the basis of practical reasoning from what is intended. Choice and intention are differently affected by standards of good reasoning, on the one hand, and concerns with further reasoning and action, on the other.

For good reason, Cohen and Levesque have as a title for their article *intention is choice with commitment*. Choice presupposes an availability of various alternatives. Intention is a choice which presupposes a measure of commitment.

3.2.8 Plans

A plan may be seen as a *recipe-for-action*. In that case, a plan is a sequence of (basic) actions that will achieve some stated goal, given some initial situation. A plan may also be seen as a complex mental attitude, as a state of mind. In that case, a plan is an abstract structure in the mind. Pollack [37] argues that only if plans are seen as structured collections of beliefs and intentions, it is possible to reason about invalid plans. Studying plan inference in this way, discrepancies are allowed between an agent's own beliefs and the beliefs that he ascribes to an actor when the agent thinks the actor has some plan. Pollack provides us the following definition, in which the connection between plans and intentions stands out clearly [37, p.89].

Definition 3.2.8.1 An agent m has a plan do α that consists in doing some set of (basic) actions $a_0, \ldots, a_n$, provided that

1. m believes that he can execute each action a_i,
2. m believes that executing $a_0, \ldots, a_n$ will entail the performance of α,
3. m believes that each action a_i plays a role in his plan,
4. m intends to execute each action a_i,
5. m intends to execute $a_0, \ldots, a_n$ as a way of doing α, and
6. m intends each action a_i to play a role in his plan.

Plans are thus a derivative of intentions and beliefs and, therefore, I shall not pay further attention to the concept of plans in this paper.

3.3 Towards a Comparison of some Formalizations of Motivational Attitudes

In this section, I will try to produce order out of the chaos in the discussion on motivational attitudes. In the first subsection, I will collect some undesired properties. What consequences of a theory of rational agency concerning motivational attitudes have to be avoided? In order to use these properties as test criteria for a comparison of agent theories in Chapter 4, I will provide formal definitions. In the second subsection, I will discuss some desired properties.

3.3.1 Undesired Properties

In this subsection an account of some undesired properties then. Since some authors use goals instead of intentions as the main operator concerning motivational attitudes, I will use MA to denote the main motivational attitude ($GOAL$ or $INTEND$).

C1 $\models \varphi \Rightarrow \models MA(\varphi)$ (necessitation rule)

It would be undesirable for formulae which represent motivational attitudes that they are validated by the necessitation rule (N-axiom). This is sometimes called the transference problem [40]. The problem of transference is illustrated by the example of an agent who intends necessary facts like the rising of the sun in the east tomorrow morning. If the necessitation rule would be validated, then intentions would not be future-directed (Dim-2).

C2 $\models (\varphi \rightarrow \psi) \Rightarrow \models (MA(\varphi) \rightarrow MA(\psi))$ (closure under logical implication)

It would be undesirable for motivational attitudes that they are closed under logical implication. Closure properties are in general undesirable for human agents (Dim-6). These problems resemble the undesired properties of logical omniscience as discussed in the previous chapter. An agent may not have realized the appropriate connection or may have realized it, but does not prefer it, nevertheless. For example, you may intend to be operated on, but even though (let us stipulate) that entails spending a day in a hospital, you may not intend spending a day in a hospital [48, p.58].

C3 $\models (MA(\varphi) \wedge MA(\varphi \rightarrow \psi)) \rightarrow MA(\psi)$ (closure under modus ponens)

It would be undesirable for motivational attitudes that they are closed under modus ponens (K-axiom). If an agent intends to go to Cologne and he intends to go to the museum Ludwig provided that he goes to Cologne, then the agent does not have to intend to go to the museum Ludwig at any price. May be, the agent will never reach Cologne [7, p.141]. To deal with this problem of conditional motivational attitudes, Buekens [7] distinguishes between external and internal conditions.

C4 $\models BEL(\varphi) \rightarrow MA(\varphi)$ (beliefs imply goals/intentions)

Further, is is not desirable that an agent has to intend everything he believes to be true. This is a weakening of the necessitation rule for motivational attitudes.

C5 $\models (MA(\varphi) \wedge BEL(\varphi \rightarrow \psi)) \rightarrow MA(\psi)$ (closure under expected consequences)

C6 $\models (MA(\varphi) \wedge BEL(\Box(\varphi \rightarrow \psi))) \rightarrow MA(\psi)$ (closure under necessary consequences)

It is not desirable that an agent has to intend all expected or necessary consequences of his intentions. Although beliefs and intentions are distinct concepts, they are, however,

related to each other as stated above (Dim-13). One of the most discussed problems concerning the relation between intentions and belief is the *side-effect problem* (Dim-7). To illustrate this problem, Bratman gives the example of a strategic bomber who intends to bomb a munition plant, believes that this will cause the adjacent school to blow up, but nevertheless does not to intend to blow up the school [5, p.139]. In discussing the side-effect problem several authors distinguish between *expected* consequences and consequences that are believed to be *necessary*.

C7 $\models MA(\varphi) \rightarrow MA(\varphi \vee \psi)$ (unrestricted weakening)

It is not desirable that an agent has to intend a disjunction provided he intends one of the disjuncts. This property is a special case of the closure under logical implication and is called *unrestricted weakening*. That this property is undesirable is shown by the example of an agent intending itself to be painted green, without intending being green or being crushed under a steam roller [34, p.155].

C8 $(\exists \mathcal{M}, \varphi \bullet \mathcal{M} \models MA(\varphi) \wedge BEL(\neg \Diamond \varphi))$, (inconsistency between beliefs and motivational attitudes)

It must not be the case that an agent intends a condition and at the same time he does not believe in the possibility of achieving that condition. This property is sometimes referred to as *intention-belief-inconsistency* and forms one of the two components of the *asymmetry thesis* [40] (Dim-11).

C9 $\models MA(\varphi) \rightarrow BEL(\varphi)$ (goals/intentions imply beliefs)

Intending a condition may not imply believing that condition, for intentions are future-directed (Dim-2). Stated in other words: *intention-belief-incompleteness* (Dim-12) should be allowed. A rational agent that intends to do an action, does not necessarily have to believe that he will do it (Dim-12).

C10 $(\exists \mathcal{M}, \varphi \bullet \mathcal{M} \models MA(\varphi) \wedge MA(\neg \varphi))$, (inconsistency between motivational attitudes of the same sort)

It should be excluded that someone intends a condition and its negation at the same time. An agent's intentions are assumed to be consistent (Dim-5). An agent can not intend to go to one of two bookstores and the other at the same time [4, p.22].

3.3.2 Desired Properties

On the other side, it would be desirable for motivational attitudes, that they can be combined, that they are (believed to be) satisfiable in some future state of the world (Dim-4), that they are closed under means (Dim-8), and that they are causes of actions (Dim-10).

Because intentions have to be consistent, it should be possible to combine intending φ and intending ψ into intending $\varphi \wedge \psi$. Bratman [4] criticizes Davidson [13] on his weak conception of the role of future intentions in further practical reasoning. "Rational intentions should be *agglomerative*. If at one and the same time I rationally intend to A and rationally intend to B then it should be both possible and rational for me, at the same time, to intend A and B." [4, p.22] Intentions perform a co-ordinating role and therefore it is needed to settle in advance on one of several options judged equally desirable.

The (belief of the) satisfiability of an intended condition can only be realized by imposing constraints on the model. Because intentions are not reducible to beliefs and desires [5], the concepts have to be connected in another way, for example, by imposing a constraint.

The closure under means has to do with the problem of intention adoption. Most theories on rational agents and motivational attitudes do not give a clue on where goals or intentions come from [21]. Some authors have recognized this problem and have asserted that intentions are part of plans. Singh [48] even defines intentions as the necessary consequences of performing a plan (strategy). Konolige and Pollack [30] state that agents often form intentions relative to pre-existing intentions. They *elaborate* their existing plans. Bratman notes that plans concerning ends embed plans concerning means and more general intentions embed more specific ones [5, p.29].

The fact that intentions are causes of action can not easily be formalized. Intentions comprise 'conation' as a component [3, pp.237ff]. An agent that intends a particular condition at least attempts to reach that condition. In my opinion, Van Linder et al. [33] meet this desideratum very well. They distinguish in their formalism between the *assertion* level and the *practition* level of commitments. The practition level contains commitments, which are recorded in an *agenda*.

3.4 Conclusions

In this chapter, I firstly summed up some some logical and intuitive properties that motivational attitudes (especially intentions) may, or may not, be taken to have. I gave priority to the concept of intentions, because they form most properly the link between belief or knowledge and actions. In the previous chapter, I characterized intentions as follows: intentions are the necessary link between someone's knowledge and someone's actions in order to reach a future state of the world. I think, I have proven this special status of intentions by comparing this notion with other motivational attitudes. Intentions (or at least *choice with a measure of commitment*) form a separate concept not reducable to the underlying concepts of desire , preference, etc.

The main results of my investigations concerning the various motivational attitudes are summarized in Table 3.1. I distinguish between three levels among the motivational attitudes. The differences between the attitudes at one level are relative small. The first level contains the intuitive motivational attitudes like desires and wishes. At the second level there are deliberated motivational attitudes like goals. For this level practical reasoning is required. The third level contains the conduct-controlling motivational attitudes like intentions. They require a measure of commitmment. Preferences are on the one hand intuitive motivational attitudes, on the other hand they require some practical reasoning because they are comparative. Therefore I put them at two levels.

In order to compare some formalizations of motivational attitudes in the next chapter, I lastly discussed the desirability of some (logical) relations between an agent's (pro-)attitudes. I provided some formal definitions of undesired properties concerning motivational attitudes. So, I can use them to check the adequacy of some agent theories to be discussed in the next chapter.

Table 3.1: Comparing the motivational attitudes

Motivational attitudes	
Intuitive: Desires, Preferences, and Wishes	They are only *potential influencers* of conduct
	They may be incompatible
	They do not necessarily move someone to act
Deliberated: Choices, Goals, Preferences, Wants, and Willing	They are more than just potential influencers of conduct
	They may not contradict each other
	They are the result of practical reasoning and involve a measure of belief
	They do not necessarily move someone to act
Conduct-controlling: Intentions	They move someone to act
	They lead to further reasoning and the formation of plans
	They have to be compatible with one's beliefs
	They have to be consistent
	They involve a measure of commitment

Chapter 4

Comparison of some Agent Theories

In this chapter I will look to some proposals of agent theories. I will compare the following theories: Cohen and Levesque [9], Rao and Georgeff [41], Konolige and Pollack [30], Singh [48], Huang et al. [27], and Van Linder et al. [33]. I recall that throughout this chapter I will make use of the definitions in Chapter 2 without explicit reference. Therefore, the original notation has to be adapted sometimes.

4.1 Cohen and Levesque

Following Bratman's philosophical work [5] Cohen and Levesque set out to specify the rational balance among beliefs, goals, actions, and intentions. The theory is expressed in a logic whose model theory is based on a possible-worlds semantics. They propose a logic with four primary modal operators, namely, BEL, $GOAL$, $HAPPENS$, and $DONE$. Each world σ is modeled as a linear sequence of events, similar to linear-time temporal models. Intentions are modeled as chosen goals that will be kept at least as long as certain conditions hold.

4.1.1 The Formal Framework

In this subsection I provide a formal definition of the formal framework of Cohen and Levesque. In their first-order language $\mathcal{L}$ events appear, denoted by $e, e', \ldots$ The formula $e \leq e'$ for event sequences e and e' denotes the fact that e is an initial subsequence of e'. Action expressions are built from variables ranging over sequences of events using the constructs of dynamic logic.

Definition 4.1.1.1 Syntactic rules

1. If a and b are action expressions, then $(a = b)$ and $(a \leq b) \in \mathcal{L}$.
2. If $m \in \mathcal{A}$ and $\varphi \in \mathcal{L}$, then $(BEL\ m\ \varphi)$ and $(GOAL\ m\ \varphi) \in \mathcal{L}$.
3. If $m \in \mathcal{A}$ and a is an action expression, then $(AGT\ m\ a)$, $(HAPPENS\ a)$, $(HAPPENS\ m\ a)$, $(DONE\ a)$, and $(DONE\ m\ a) \in \mathcal{L}$.

Definition 4.1.1.2 Models for interpretation
$\mathcal{M} = \langle \Theta, \mathcal{A}, E, Agt, W, B, G, V \rangle$ is a model for interpretation of formulae of $\mathcal{L}$ if

1. Θ is a set of things,
2. $\mathcal{A}$ is a set of agents,
3. E is a set of primitive event types ($D = \Theta \cup \mathcal{A} \cup E$ is used as the domain of quantification),
4. $Agt : E \mapsto \mathcal{A}$ is a function that assigns to each event type a single agent,
5. $W = \{\sigma | \sigma : Z \mapsto E\}$ is a set of functions from the integers (Z) into the set of primitive event types (W is the set of possible worlds; Z acts as a time axis.),
6. $B, G \subseteq W \times \mathcal{A} \times Z \times W$ are accessibility relations that relate to each agent at a certain moment in a given world a set of worlds that are compatible with his beliefs (B) and goals (G), and
7. V is an interpretation function defined on predicates and variables.

Definition 4.1.1.3 Semantic rules
In the following $[\![\ ]\!]$ maps action expressions to (unique) event sequences. Furthermore, $begins(e)$ and $ends(e)$ are auxiliary functions that indicate the begin time and the end time of event sequence e. It may be assumed that $begins(e) \preceq ends(e)$.

1. $\mathcal{M}, \sigma, t \models a = b$ iff $[\![a]\!] = [\![b]\!]$
2. $\mathcal{M}, \sigma, t \models a \leq b$ iff $(\exists c \bullet [\![a; c]\!] = [\![b]\!])$
3. $\mathcal{M}, \sigma, t \models (BEL\ m\ \varphi)$ iff $(\forall \tau \bullet B(\sigma, m, t, \tau) \Rightarrow \mathcal{M}, \tau, t \models \varphi)$
4. $\mathcal{M}, \sigma, t \models (GOAL\ m\ \varphi)$ iff $(\forall \tau \bullet G(\sigma, m, t, \tau) \Rightarrow \mathcal{M}, \tau, t \models \varphi)$
5. $\mathcal{M}, \sigma, t \models (AGT\ m\ a)$ iff $Agt([\![a]\!]) = m$
6. $\mathcal{M}, \sigma, t \models (HAPPENS\ a)$ iff $begins([\![a]\!]) = t$
7. $\mathcal{M}, \sigma, t \models (DONE\ a)$ iff $ends([\![a]\!]) = t$
8. $\mathcal{M}, \sigma, t \models (HAPPENS\ m\ a)$ iff $\mathcal{M}, \sigma, t \models (HAPPENS\ a) \wedge (AGT\ m\ a)$
9. $\mathcal{M}, \sigma, t \models (DONE\ m\ a)$ iff $\mathcal{M}, \sigma, t \models (DONE\ a) \wedge (AGT\ m\ a)$

Convention 4.1.1.4 Abbreviation
$\Diamond\varphi \stackrel{def}{=} (\exists a \bullet (HAPPENS a; \varphi?))$

Because this definition of $\Diamond\varphi$ resembles the semantics of the earlier defined future-operator $F\varphi$, I will use the latter to avoid confusion.

Constraint 4.1.1.5 Constraints on the model

1. The relation B is euclidean, transitive, and serial.
2. The relation G is serial.
3. If $G(\sigma, m, t, \tau)$, then $B(\sigma, m, t, \tau)$.

The first constraints guarantee consistency. The last one is the realism constraint. The worlds that are consistent with what the agent has chosen are not ruled out by his beliefs. The constraint ensures that an agent does not want the opposite of what he believes to be unchangeable.

Definition 4.1.1.6 Auxiliary definitions

1. $(LATER\ \varphi) \stackrel{\text{def}}{=} \neg\varphi \wedge \Diamond\varphi$
2. $(BEFORE\ \varphi\ \psi) \stackrel{\text{def}}{=} (\forall b \bullet (HAPPENS\ b; \psi?) \rightarrow (\exists a \bullet (a \leq b) \wedge (HAPPENS\ a; \varphi?)))$

4.1.2 Motivational Attitudes

The formula $(BEL\ m\ \varphi)$ is true iff φ is true in all belief-accessible worlds relative to m. The worlds the agent thinks are possible do not necessarily include the real world. On the other side, an agent believes all the inevitable true formulae: If $\models \varphi$, then $(BEL\ m\ \varphi)$.

Definition 4.1.2.1 Knowledge and competence

1. $K_m(\varphi) \stackrel{\text{def}}{=} \varphi \wedge (BEL\ m\ \varphi)$
2. $(COMPETENT\ m\ \varphi) \stackrel{\text{def}}{=} (BEL\ m\ \varphi) \rightarrow \varphi$

Knowledge is, as usual, identified with true belief. Agents are assumed to be *competent* with respect to the primitive actions they have done. Further, if an agent thinks he is about to do something, then there must be some initial sequence that he believes he is going to do next.

The formula $(GOAL\ m\ \varphi)$ is meant to be read as φ is true in all worlds, accessible from the current world, that are compatible with m's goals. Since agents choose entire worlds, they choose the (logically and physically) necessary consequences of their goals. Goals are consistent and what is implicit in someone's goals is closed under consequence. As with the beliefs of an agent, we have the necessitation property for goals: If $\models \varphi$, then $\models (GOAL\ m\ \varphi)$. *Achievement goals* are those goals the agent believes to be currently false.

Definition 4.1.2.2 Achievement goals

$(A'GOAL\ m\ \varphi) \stackrel{\text{def}}{=} (GOAL\ m\ (LATER\ \varphi)) \wedge (BEL\ m\ \neg\varphi)$

It is assumed that agents eventually drop all achievement goals. Therefore, the following constraint is imposed on the formal model.

Constraint 4.1.2.3 No persistence/deferral forever

$\models \mathsf{F}\neg(GOAL\ m\ (LATER\ \varphi))$

As a constraint it is said that worlds compatible with an agent's goals must be included in those compatible with his beliefs. If φ holds in all belief-accessible worlds relative to an agent m, then φ must also be true in the relevant goal-accessible worlds. Therefore, the following lemma can be stated.

Lemma 4.1.2.4 $\models (BEL\ m\ \varphi) \rightarrow (GOAL\ m\ \varphi)$

As a corollary of the above lemma the following formulae can be asserted. The first one states that an agent's beliefs and goals 'line up' with respect to his own primitive actions that happen next. From the second it can be learned that goals are closed under beliefs.

Corollary 4.1.2.5 Implications of the previous lemma

1. $\models (\forall m, e \bullet (BEL\ m\ (HAPPENS\ m\ e)) \rightarrow (GOAL\ m\ (HAPPENS\ m\ e)))$
2. $\models (GOAL\ m\ \varphi) \wedge (BEL\ m\ (\varphi \rightarrow \psi)) \rightarrow (GOAL\ m\ \psi)$

To capture a grade of commitment that an agent might have towards his goals, a *persistent goal* is defined as an achievement goal that the agent will not give up until he thinks it has been satisfied, or until he thinks it will never be true.

Definition 4.1.2.6 Persistent goal

$(P'GOAL\ m\ \varphi) \stackrel{\text{def}}{=} (A'GOAL\ m\ \varphi) \wedge (BEFORE\ \chi\ \neg(GOAL\ m\ (LATER\ \varphi)))$,
where $\chi \equiv (BEL\ m\ \varphi) \vee (BEL\ m\ \mathsf{G}\neg\varphi)$

The logic of $P'GOAL$ is weaker than one might expect. Unlike $GOAL$, $P'GOAL$ does not distribute over conjunction or disjunction, and it is closed only under logical equivalence. Cohen and Levesque define also *persistent relativized goals*. The notation is as follows: $(PR'GOAL\ m\ \varphi\ \psi)$. In that case there is another reason to drop one's goals. A necessary condition for giving up a $PR'GOAL$ is that the agent believes it is satisfied, or believes it is unachievable, *or* believes $\neg\psi$.

Definition 4.1.2.7 Persistent relativized goal

$(PR'GOAL\ m\ \varphi\ \psi) \stackrel{\text{def}}{=} (A'GOAL\ m\ \varphi) \wedge (BEFORE\ \chi\ \neg(GOAL\ m\ (LATER\ \varphi)))$,
where $\chi \equiv (BEL\ m\ \varphi) \vee (BEL\ m\ \mathsf{G}\neg\varphi) \vee (BEL\ m\ \neg\psi)$

Further, Cohen and Levesque define two forms for $INTEND$, depending on whether the argument is an action or a proposition.

Definition 4.1.2.8 Intentions towards actions and propositions

1. $(INTEND_1\ m\ a) \stackrel{\text{def}}{=} (P'GOAL\ m\ (DONE\ m\ (BEL\ m\ (HAPPENS\ a))?; a))$,
   where a is any action expression
2. $(INTEND_2\ m\ \varphi) \stackrel{\text{def}}{=} (P'GOAL\ m\ (\exists e \bullet (DONE\ m\ \chi?; e; \varphi?)))$, where
   $\chi \equiv ((BEL\ m\ (\exists e' \bullet (HAPPENS\ m\ e'; \varphi?))) \wedge \neg(GOAL\ m\ \neg(HAPPENS\ m\ e; \varphi?)))$

An intentions towards an action is defined as a persistent goal to having done an action a deliberately. The agent will not be committed to doing something accidently or unknowingly. An intentions towards a proposition is defined as a persistent goal to having done some sequence of events e himself, after which φ holds. Prior to doing e to bring about φ, the agent believes that he is about to do something (event sequence e') bringing about φ, and the agent does not have as a goal not doing e to bring about φ.

4.1.3 Evaluation

Singh [47] criticizes the fact that the assumption that agents eventually drop their goals, does not involve actions and abilities of agents. As a result, the policies for goal revision do not address three important properties: i) ability of an agent to act, ii) that the agent actually acts for the goal, and iii) success or failure of an action. Intentions in this formalism do not necessarily move an agent to act. Further, the theory does not capture the essential distinction between the semantics of intentions and policies of when to update them and when not. "If an agent who intends to go to a museum learns that the bridge he planned

to drive on is closed, he might give up his original intention, rather than rent an expensive helicopter." In the formalism of Cohen and Levesque, it has to be stated in advance the conditions under which a particular intention may be given up. However, the entire set of possible exceptions cannot be specified in advance. As an overall judgement Singh asserts the following: "The nesting of the definitions makes Cohen and Levesque's theory the most complicated of the works on intentions." [48, p.76]

According to Haddadi [23], the resulting theory does not necessarily characterize introspective agents. Therefore, the theory cannot be embedded straightforwardly in the operational model of an autonomous agent. Another point of criticism is that, in contrast to the title of their work (intention is choice with commitment), the theory does not explicitly capture the notion of choice.

Konolige and Pollack [30] argue that, in the formalism of Cohen and Levesque, an agent who always believes that $\varphi \rightarrow \psi$ is always true will incur the *side-effect problem* when intending φ, i.e., the agent has to intend ψ unintentionally. Formally: $(INTEND\ m\ \varphi) \wedge \mathsf{G}(BEL\ m\ \mathsf{G}(\varphi \rightarrow \psi)) \rightarrow (INTEND\ m\ \psi)$.

Besides, the closure under logical implication (C2) and unrestricted weakening (C7) are validated. Cohen and Levesque defend their formalism by stipulating that an agent can believe the side-effect (ψ) already holds. Because an agent can only intend (have as a persistent goal) propositions that do not hold already, agents will not be forced to adopt a side-effect as an intention. However, this argument seems not very convincing to me. To come back to one of the examples of the previous chapter, an agent intending to be operated on does surely not believe that he is spending a day in a hospital. Fortunately, Cohen and Levesque admit that closure properties are avoided "because of the wrong reasons." [9, p.238]

4.2 Rao and Georgeff

Rao and Georgeff provide an adapted formalism. The most components are due to Cohen and Levesque [9]. Two of the main differences are the following. Instead of linear time space, Rao and Georgeff use *time trees*, which are temporal structures with a branching time future and a linear past. Further, the authors define, in opposition to Cohen and Levesque, intentions as basic entities, irreducible to the other basic attitudes of belief and desire (goal).

4.2.1 The Formal Framework

The formal theory that Rao and Georgeff propose is based on CTL^*. This logic is extended to a first-order possible-worlds framework by introducing modal operators for beliefs, goals, and intentions. In the logic two types of formulae occur: state formulae and path formulae. $\mathcal{L}_s$ denotes the state formulae, $\mathcal{L}_p$ denotes the path formulae. It may be clear that $\mathcal{L}_s \subseteq \mathcal{L}_p$

Definition 4.2.1.1 Syntactic rules

1. If e is an event type, then $succeeds(e)$, $fails(e)$, $does(e)$, $succeeded(e)$, $failed(e)$, and $done(e) \in \mathcal{L}_s$
2. If $\varphi \in \mathcal{L}_s$, then $BEL(\varphi)$, $GOAL(\varphi)$ and $INTEND(\varphi) \in \mathcal{L}_s$

Furthermore, the authors use time-operators, like $\mathsf{E}\varphi$, $\mathsf{A}\varphi$, $\mathsf{F}\varphi$, $\mathsf{G}\varphi$, $\varphi\mathsf{U}\psi$, $\bigcirc\varphi$ as they are defined in Chapter 2.

Remark 4.2.1.2 Rao and Georgeff use in their notation $\Diamond\varphi$ instead of $\mathsf{F}\varphi$, $\Box\varphi$ instead of $\mathsf{G}\varphi$, $optional(\varphi)$ instead of $\mathsf{E}\varphi$, and $inevitable(\varphi)$ instead of $\mathsf{A}\varphi$.

Belief is modeled in the conventional way. That is, in each situation a set of *belief-accessible* worlds is associated; intuitively, those worlds that the agent believes to be possible. Similar to belief-accessible worlds, we have for each situation a set of *goal-accessible* worlds. They represent the goals of the agent. Goals are chosen desires of the agent that are consistent. Moreover, the agent should belief that the goals are achievable. Intentions are represented by sets of *intention-accessible* worlds. These worlds are ones that the agent has committed to attempt to realize. An agent can only intend some course of action if it is one of his goals. An agent moves from a belief-accessible world to a goal-accessible world by *desiring* future paths, and from a goal-accessible world to an intention-accessible world by *committing* to certain desired future paths.

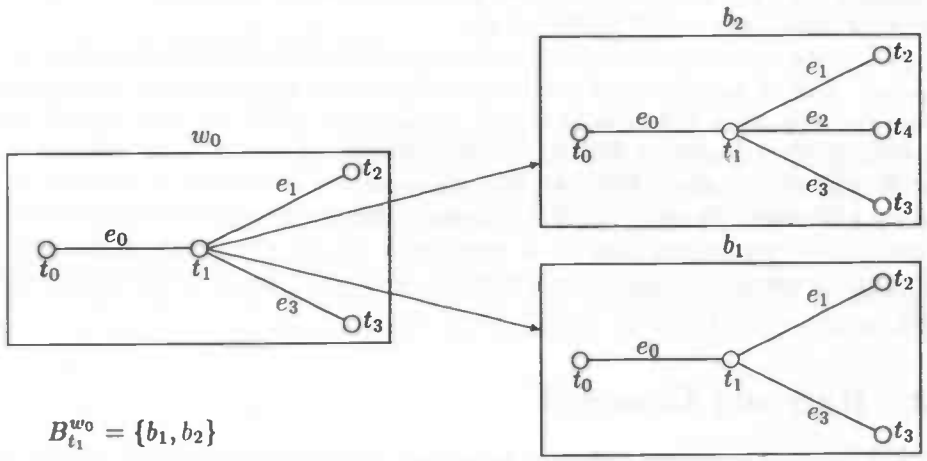

Figure 4.1: Worlds as time trees

Definition 4.2.1.3 Models for interpretation
$\mathcal{M} = \langle W, E, \mathbf{T}, \prec, U, B, G, I, V \rangle$ is a model for interpretation of formulae of $\mathcal{L}$ if

1. W is a set of worlds,
2. E is a set of primitive event types,
3. $\mathbf{T}$ is a set of time points,
4. $\prec$ is a binary relation on $\mathbf{T}$,
5. U is the universe of discourse,
6. $B : W \times \mathbf{T} \mapsto \wp(W)$ is a function that maps the agent's current situation to his belief-accessible worlds,
7. $G : W \times \mathbf{T} \mapsto \wp(W)$ is a function that maps the agent's current situation to his goal-accessible worlds,

8. $I : W \times \mathbf{T} \mapsto \wp(W)$ is a function that maps the agent's current situation to his intention-accessible worlds, and
9. V is a mapping of first-order entities to elements in U for any given world and time point.

Notation 4.2.1.4 A situation is a world, say w, at a particular time point, say t, and is denoted by w_t. B_t^w denotes the set of worlds belief-accessible from world w at time t. G_t^w and I_t^w are similarly defined.

Definition 4.2.1.5 Temporal structure

1. Each world w of W, called a *time tree*, is a tuple $\langle \mathbf{T}_w, \mathcal{A}_w, \mathcal{S}_w, \mathcal{F}_w \rangle$.
2. A world w' is a *sub-world* of the world w, denoted by $w' \sqsubseteq w$, iff
   (a) $\mathbf{T}_{w'} \subseteq \mathbf{T}_w$
   (b) $(\forall t \in \mathbf{T}_{w'} \bullet V(P, w', t) = V(P, w, t))$, where P is a predicate symbol
   (c) $(\forall t \in \mathbf{T}_{w'} \bullet \mathcal{R}_t^w = \mathcal{R}_t^{w'})$, where $\mathcal{R}$ is used to denote one of B, G, and I
   (d) $\mathcal{A}_{w'}$, $\mathcal{S}_{w'}$ and $\mathcal{F}_{w'}$ are equal to $\mathcal{A}_w$, $\mathcal{S}_w$ and $\mathcal{F}_w$ restricted to time points in $\mathbf{T}_{w'}$

$\mathbf{T}_w \subseteq \mathbf{T}$ is a set of time points in the world w and $\mathcal{A}_w$ is the same as $\prec$, restricted to time points in $\mathbf{T}_w$. A *fullpath* in a world w, denoted by $(w_{t_0}, w_{t_1}, \ldots)$, is an infinite sequence of time points such that $(\forall i \bullet (t_i, t_{i+1}) \in \mathcal{A}_w)$. The arc functions $\mathcal{S}_w$ and $\mathcal{F}_w$ map adjacent time points to events in E. Intuitively, for any two adjacent time points for which the arc function $\mathcal{S}_w$ ($\mathcal{F}_w$) is defined, its value represents the event that successfully occurred (failed) between those time points. It is required that if $\mathcal{S}_w(t_i, t_j) = \mathcal{S}_w(t_i, t_k)$, then $t_j = t_k$ and similarly for $\mathcal{F}_w$.

Constraint 4.2.1.6 Constraints on the model

For every $w \in W$ and for every $t \in \mathbf{T}$ the following constraints are imposed on the models defined above.

1. $(\forall w' \in B_t^w \bullet (\exists w'' \in G_t^w \bullet w'' \sqsubseteq w'))$
2. $(\forall w' \in G_t^w \bullet (\exists w'' \in I_t^w \bullet w'' \sqsubseteq w'))$
3. $(\forall w' \in B_t^w, w'' \in I_t^w \bullet w'' \in B_t^{w'})$
4. $(\forall w' \in B_t^w, w'' \in G_t^w \bullet w'' \in B_t^{w'})$
5. $(\forall w' \in G_t^w, w'' \in I_t^w \bullet w'' \in G_t^{w'})$

In words: for every belief-accessible world w' there exists a goal-accessible world w'' that is a subworld of w'. For every goal-accessible world w' there exists an intention-accessible world w'' that is a subworld of w'. For every belief-accessible world w' and every intention-accessible world w'' it holds that w'' is belief-accessible from w'. For every belief-accessible world w' and every goal-accessible world w'' it holds that w'' is belief-accessible from w'. For every goal-accessible world w' and every intention-accessible world w'' it holds that w'' is goal-accessible from w'.

Definition 4.2.1.7 Semantic rules

The truth-value of state formulae is evaluated relative to a world, w, and a moment, t. For the semantic rules of the time-operators, the reader is referred to Chapter 2.

1. $\mathcal{M}, w_{t_1} \models succeeded(e)$ iff $(\exists t_0 \bullet \mathcal{S}_w(t_0, t_1) = e)$
2. $\mathcal{M}, w_{t_1} \models failed(e)$ iff $(\exists t_0 \bullet \mathcal{F}_w(t_0, t_1) = e)$
3. $\mathcal{M}, w_t \models BEL(\varphi)$ iff $(\forall w' \in B_t^w \bullet \mathcal{M}, w'_t \models \varphi)$
4. $\mathcal{M}, w_t \models GOAL(\varphi)$ iff $(\forall w' \in G_t^w \bullet \mathcal{M}, w'_t \models \varphi)$
5. $\mathcal{M}, w_t \models INTEND(\varphi)$ iff $(\forall w' \in I_t^w \bullet \mathcal{M}, w'_t \models \varphi)$

Definition 4.2.1.8 The semantic rules for $done(e)$, $succeeds(e)$, $fails(e)$, and $does(e)$ can be defined with the help of the above defined semantic rules.

1. $done(e) \stackrel{\text{def}}{=} succeeded(e) \vee failed(e)$
2. $succeeds(e) \stackrel{\text{def}}{=} \mathsf{A} \bigcirc (succeeded(e))$
3. $fails(e) \stackrel{\text{def}}{=} \mathsf{A} \bigcirc (failed(e))$
4. $does(e) \stackrel{\text{def}}{=} \mathsf{A} \bigcirc (done(e))$

Rao and Georgeff present the following set of axioms. Some of these axioms are realized in the formal framework by the above defined constraints. In this way the desired interrelationships among an agent's beliefs, goals, and intentions are captured.

Definition 4.2.1.9 A set of axioms

In this definition α is a well-formed formula that contain no positive occurrence of A outside the scope of the modal operators BEL, $GOAL$, or $INTEND$. φ is an arbitrary formula from $\mathcal{L}$.

1. $GOAL(\alpha) \rightarrow BEL(\alpha)$
2. $INTEND(\alpha) \rightarrow GOAL(\alpha)$
3. $INTEND(does(e)) \rightarrow does(e)$
4. $INTEND(\varphi) \rightarrow BEL(INTEND(\varphi))$
5. $GOAL(\varphi) \rightarrow BEL(GOAL(\varphi))$
6. $INTEND(\varphi) \rightarrow GOAL(INTEND(\varphi))$
7. $done(e) \rightarrow BEL(done(e))$
8. $INTEND(\varphi) \rightarrow \mathsf{AF}(\neg INTEND(\varphi))$

4.2.2 Commitment Strategies

A commitment strategy formalizes the process of intention maintenance and revision. Rao and Georgeff describe three different commitment strategies and present accompanying axioms. A *blindly* committed agent maintains his intentions until he actually believes that he has achieved them. Relaxing this requirement, they define *single-minded* commitment, in which the agent maintains his intentions as long as he believes that they are still options. As long as he believes his intentions to be achievable, a single-minded agent will not drop his intentions and thus is committed to his goals. This requirement can also be relaxed. An *open-minded* agent maintains his intentions as long as these intentions are still his goals. The following axioms formalize the three commitment strategies successively.

Definition 4.2.2.1 Axioms for commitment strategies

1. $INTEND(\mathsf{AF}\varphi) \Rightarrow \mathsf{A}(INTEND(\mathsf{AF}\varphi)\mathsf{U}BEL(\varphi))$
2. $INTEND(\mathsf{AF}\varphi) \Rightarrow \mathsf{A}(INTEND(\mathsf{AF}\varphi)\mathsf{U}BEL(\varphi) \vee \neg BEL(\mathsf{EF}\varphi))$
3. $INTEND(\mathsf{AF}\varphi) \Rightarrow \mathsf{A}(INTEND(\mathsf{AF}\varphi)\mathsf{U}BEL(\varphi) \vee \neg GOAL(\mathsf{EF}\varphi))$

A blindly committed agent will not give up his intentions until he comes to believe that he has achieved his intentions. A single-minded agent reaches an identical conclusion only if he continues to believe, until the time he believes he has realized his intentions, that the intended ends remains an option. Similarly, an open-minded agent will eventually believe he has achieved his intentions provided he maintains these intentions as goals until they are believed to have been achieved.

4.2.3 Evaluation

Rao and Georgeff distinguish between *successful actions* and *failures*, but they do not clarify the meaning of primitive action. Actions and events are not full part of their formalism (in opposition to the theories of Cohen and Levesque, and Singh). Maybe that is the reason of not defining constraints in which actions are linked with the mental attitudes.

Rao and Georgeff argue that with their *strong realism* (at least one intended world is a belief world) intentions can be made side-effect free. However, Konolige and Pollack argue that this sort of realism seems inherently less desirable than the realism presented by Cohen and Levesque (how is it possible for an agent to intend worlds he does not believe possible?), and it is still not fully side-effect free, since it is closed under conjunctions and abstractions [30, p.391].

All necessary truths are intended by the agent, because they are true in all the intended worlds. So, this formalism validates the necessitation rule (C1). The theory also validates the closure under logical implication (C2), the closure under modus ponens (C3), and unrestricted weakening (C7). Rao and Georgeff present as an axiom of their theory that what one intends entails what one believes. So (C9) is also validated. However, the theory does not validate the closure under beliefs and, therefore, avoids the side-effect problem.

In [40] Rao and Georgeff propose *weak realism* in order to avoid intending undesired side-effects. Their weak-realism constraint requires that there be at least one world common to belief- and goal-accessible worlds, and similarly for belief- and intention-accessible worlds and goal- and intention-accessible worlds.

4.3 Konolige and Pollack

In their article [30], Konolige and Pollack argue that normal modal logics are not an appropriate representation for intention: The semantic rule for normal modal operators leads to the confusion of an intention to do φ with an intention to do any logical consequence of φ, called the side-effect problem. Moreover, normal modal logics do not provide a means of relating intentions to one another.

4.3.1 The Formal Framework

The model of Konolige and Pollack has two components: 1) A set of possible worlds, W, that represent possible future courses of events; 2) A set of cognitive structures, $\langle W, \Sigma, I \rangle$, that represent the mental state components of an agent. The formalism is not based on

an explicit temporal structure. Each possible world is a complete history, specifying states of the world at all instants of time. Konolige and Pollack only concentrated on the static relation between intention and belief. The evaluation point for statements is the same in all worlds.

Definition 4.3.1.1 Syntactic rules

1. If $\varphi \in \mathcal{L}$ then $\Box\varphi$ and $\Diamond\varphi \in \mathcal{L}$
2. If $\varphi \in \mathcal{L}$ then $BEL(\varphi)$ and $INTEND(\varphi) \in \mathcal{L}$

The accessibility relation for the modal operators $\Box$ and $\Diamond$ is the universal relation. $\Diamond\varphi$ says that there is a world $w \in W$ for which φ is true.

Definition 4.3.1.2 Models for interpretation
$\mathcal{M} = \langle W, \Sigma, I \rangle$ (also called a *cognitive structure for an agent*) is a model for interpretation of formulae of $\mathcal{L}$ if

1. W is a set of possible worlds,
2. Σ is a set of belief-accessible worlds ($\Sigma \subseteq W$), and
3. I is a set of sentences of $\mathcal{L}$ which represent the intentions of the agent.

Definition 4.3.1.3 Scenarios
If $\varphi \in \mathcal{L}$ then $W_\varphi = \{w \in W | \mathcal{M}, w \models \varphi\}$. W_φ corresponds to the subset of W that makes φ true, and is called a *scenario* for φ.

Definition 4.3.1.4 Semantic rules

1. $\mathcal{M} \models \Diamond\varphi$ iff $(\exists w' \in W \bullet M, w \models \varphi)$
2. $\mathcal{M} \models BEL(\varphi)$ iff $(\Sigma \subseteq W_\varphi)$
3. $\mathcal{M} \models INTEND(\varphi)$ iff $(\exists \psi \in I \bullet W_\varphi = W_\psi)$
4. $\mathcal{M} \models INTEND^*(\varphi)$ iff $(\exists J \subseteq I \bullet W_\varphi = W_J)$

$INTEND(\varphi)$ is used to refer to the *primary intentions* of the agent contained in the set I. Primary intentions do not depend on any other intentions that the agent currently has. $INTEND^*(\varphi)$ is used to refer to complex intentions, called *conjoined intentions*. In the semantic rule for $INTEND(\varphi)$ equality is used. As a consequence, intentions are only closed under logical equivalence. In the alternative reading of intentions, $INTEND^*$, φ is intended if it is the intersection of worlds of some set of primary intentions. The following axioms characterize the properties of $INTEND^*$.

Lemma 4.3.1.5 Properties of intentions

1. $INTEND(\varphi) \Rightarrow INTEND^*(\varphi)$
2. $INTEND^*(\varphi) \wedge INTEND^*(\psi) \Rightarrow INTEND^*(\varphi \wedge \psi)$

Constraint 4.3.1.6 Constraints on the model

1. $(\exists w \in \Sigma \bullet (\forall \varphi \in I \bullet w \in W_\varphi))$

2. $(\forall \varphi \in I \bullet (\exists w \in \Sigma \bullet w \notin W_\varphi))$

In words: there is a belief-accessible world in which every primary intention is realized. So, the intentions of an agent are not ruled out by his beliefs. A rational agent will not form intentions that he does not believe to be realizable in connection with his other intentions. In the second constraint, it is stated that for every primary intention there exists a belief-accessible world in which that intention is not realized. Intentions should be nontrivial, in the sense that the agent intending φ should not believe that φ will occur without the intervening action of the agent.

Lemma 4.3.1.7 The following sentences are valid in all models.

1. $\neg INTEND(\varphi \wedge \neg\varphi)$ (consistency)
2. $INTEND(\varphi) \wedge INTEND(\psi) \Rightarrow \Diamond(\varphi \wedge \psi)$ (joint consistency)
3. $INTEND^*(\varphi) \Rightarrow \Diamond\varphi$
4. $INTEND^*(\varphi) \Rightarrow BEL(\Diamond\varphi)$ (realism)
5. $INTEND(\varphi) \Rightarrow \neg BEL(\neg\varphi)$ (epistemic consistency)
6. $INTEND(\varphi) \wedge INTEND(\psi) \Rightarrow \neg BEL(\neg(\varphi \wedge \psi))$ (joint epistemic consistency)
7. $INTEND^*(\varphi) \Rightarrow \neg BEL(\neg\varphi)$
8. $INTEND(\varphi) \Rightarrow \neg BEL(\varphi) \wedge \neg BEL(\neg\varphi)$ (epistemic indeterminacy)

4.3.2 Relative Intentions

Agents often form intentions relative to pre-existing intentions or, in other words, they *elaborate* their existing plans. Konolige and Pollack introduce a graph among intentions to formalize this phenomenon. The language is extended to include a modal operator By.

Definition 4.3.2.1 Syntactic rule
If $\varphi, \psi_1, \ldots, \psi_n \in \mathcal{L}$, then $By(\varphi; \psi_1, \ldots, \psi_n) \in \mathcal{L}$.

In this formula, $\psi_1, \ldots, \psi_n$ are called an *elaboration* of φ. The models for interpretation are extended accordingly with a new component, $\longrightarrow$, which is a graph relating intentions to one another. Such extended models are called *elaborated cognitive structures*. The graphs are acyclic and rooted in the set of primary intentions.

Definition 4.3.2.2 Models for interpretation
$\mathcal{M} = \langle W, \Sigma, I, \longrightarrow \rangle$ is a model for interpretation of formulae of $\mathcal{L}$ if

1. $\mathcal{M}$ satisfies the requirements mentioned in definition 4.3.1.2, and
2. $\longrightarrow \subseteq (\wp(\mathcal{L}) \times \mathcal{L})$ is a relation that represents the means-end structure of an agent's intentions.

Definition 4.3.2.3 Semantic rule
$\mathcal{M} \models By(\varphi; \psi_1, \ldots, \psi_n)$ iff $\psi_1, \ldots, \psi_n \longrightarrow \varphi$

It would be desirable that an agent believes that his intention is realized, as soon as the elaboration of his intention is achieved. In order to ensure this, Konolige and Pollack define a constraint. They call a cognitive structure *embedded* if it satisfies the following constraint.

Constraint 4.3.2.4 Embedded cognitive structures
$\psi_1, \ldots, \psi_n \longrightarrow \varphi \Rightarrow \bigcap_{i=1}^{n} W_{\psi_i} \subseteq W_\varphi$

Lemma 4.3.2.5 Property of embedded structures
$\langle W, \Sigma, I, \longrightarrow \rangle \models By(\varphi; \psi_1, \ldots, \psi_n) \Rightarrow BEL(\psi_1 \wedge \ldots \wedge \psi_n \rightarrow \varphi)$

The embedding graph, $\longrightarrow$, represents the structure of intentions in a direct way, by means of a relation among the relevant scenarios. A normal modal logic is incapable of this, because its accessibility relation goes from a single world (rather than a scenario) to a set of possible worlds.

4.3.3 Evaluation

The set I of intentions is not closed with respect to valid consequence (side-effects, conjunctions, and abstractions). Moreover, I is not closed under an agent's beliefs. The theory of Konolige and Pollack does not validate any of the criteria mentioned in Chapter 3. However, when it comes to expressibility, the results are not very good. The main aim of their article is the static relation between intentions and belief. Haddadi [23] admits that syntactic approaches avoid the problems associated with logical omniscience, but in his view these approaches are semantically not as strong as the possible-worlds approaches.

4.4 Singh

Singh uses a formal model that is based on possible worlds, too. The possible worlds here, in the technical sense of the term, are possible moments. That is, each *moment* plays the role of a world in standard modal logic. Each moment is associated with a possible state of the world, which is identified by the conditions that hold at that moment. At each moment, environmental events, and agents' actions (possibly idle) occur. The same physical state may occur at different moments. Further, Singh uses a branching time structure. At each moment there are several branches, or *scenarios*, along which a world can progress, depending on the choices of the agents in the system. One of the scenarios beginning at a moment is identified as the real one. This is the scenario on which the world can be seen to have progressed, assuming it was in the state denoted by the given moment. The real scenario is determined by the choices of the agents and events in the environment.

Remark 4.4.0.1 While in the framework of Konolige and Pollack [30] a scenario denotes a set of worlds satisfying a given formula, in the framework of Singh a scenario denotes a path in a branching time structure. Another point of difference worthy remarking is the fact that Singh uses a continuous time structure in opposition to Cohen and Levesque [9], and Rao and Georgeff [41]. The latter use a discrete time structure.

4.4.1 The Formal Framework

The proposed formal language, $\mathcal{L}$, is based on CTL^*. $\mathcal{L}_p$ denotes the set of "scenario-formulae," which are just path formulae. $\mathcal{L}_s$ contains the state formulae. They are evaluated relative to moments. The formulae in $\mathcal{L}_p$ are evaluated relative to scenarios and moments.

Definition 4.4.1.1 Syntactic rules

1. $\mathcal{L}_s \subseteq \mathcal{L}_p$

2. If $\varphi \in \mathcal{L}_p$ then $\mathsf{A}\varphi$, $\mathsf{E}\varphi$, and $\mathsf{R}\varphi \in \mathcal{L}$
3. If $\varphi \in \mathcal{L}$ and $x \in \mathcal{X}$ then $(\bigvee x : \varphi) \in \mathcal{L}$
4. If $\varphi, \psi \in \mathcal{L}_p$ then $\varphi \mathsf{U} \psi \in \mathcal{L}_p$
5. If $\varphi \in \mathcal{L}_p$, $m \in \mathcal{A}$, and $a \in \mathcal{B}$ then $m[a]\varphi$, $m\langle a\rangle\varphi$, and $m \langle\!|a|\!\rangle \varphi \in \mathcal{L}_p$

Definition 4.4.1.2 Models for interpretation
$\mathcal{M} = \langle \mathbf{T}, \prec, [\![\,]\!], \mathbf{B}, \mathbf{R}\rangle$ is a model for interpretation of formulae of $\mathcal{L}$ if

1. $\mathbf{T}$ is a set of possible moments ordered by $\prec$,
2. $[\![\,]\!]$ is an interpretation function,
3. $\mathbf{B} : \mathcal{A} \mapsto \wp(\mathbf{T} \times \mathbf{T})$ assigns alternative moments to the agents at each moment, and
4. $\mathbf{R} : \mathbf{T} \mapsto \wp(\mathbf{T})$ assigns a scenario to each moment, which is interpreted as the real scenario at that moment.

A *scenario* at a moment is any single branch of the relation $\prec$ that begins at the given moment, and contains all moments in some linear subrelation of $\prec$. $\mathbf{S}_t$ is the set of all scenarios at moment t. $[\sigma; t, t']$ denotes a period on scenario σ from t to t'. Formally, a scenario at moment t is a set $\sigma \subseteq \mathbf{T}$ of which the following conditions hold.

Definition 4.4.1.3 Conditions for scenarios
If $\sigma \in \mathbf{S}_t$ then the following conditions hold.

1. $t \in \sigma$
2. $(\forall t, t' \in \sigma \bullet t = t' \vee t \prec t' \vee t' \prec t)$
3. $(\forall t, t' \in \sigma, t'' \in \mathbf{T} \bullet (t \prec t'' \prec t') \Rightarrow t'' \in \sigma)$
4. $(\forall t' \in \sigma, t'' \in \mathbf{T} \bullet (t' \prec t'') \Rightarrow (\exists t''' \in \sigma \bullet (t' \prec t''') \wedge \neg(t''' \prec t'')))$

Informally, this means that a scenario at moment t is a set that contains t, has a linear order, contains all the moments within that order, and can always be extended along some particular branch. The following properties for scenarios and periods can be defined.

Definition 4.4.1.4 Properties of (sets of) scenarios

1. $t \neq t' \Rightarrow \mathbf{S}_t \cap \mathbf{S}_{t'} = \emptyset$
2. $t, t' \Rightarrow (\forall \tau \in \mathbf{S}_{t'} \bullet (\exists \sigma \in \mathbf{S}_t \bullet \tau \subset \sigma))$
3. $(\forall \sigma \in \mathbf{S}_t, t' \in \sigma \bullet (\exists \tau \in \mathbf{S}_{t'} \bullet \tau \subseteq \sigma))$
4. $[\sigma; t, t'] \subseteq \tau \Rightarrow [\sigma; t, t'] = [\tau; t, t']$

The interpretation, $[\![\,]\!]$, gives the semantic content of some of the symbols of the formal language. The interpretation of an atomic proposition is the set of moments at which it is true. The interpretation of an action symbol a is, for each agent symbol m, the set of periods in the model in which an instance of a is done by m. $[\sigma; t, t'] \in [\![a]\!]^m$ means that agent m is performing action a on scenario σ from moment t to moment t'.

Singh explicitly distinguishes between *know-that* and *know-how*. He uses the former in the normal sense of knowledge. The notion of know-how has to do with the ability to perform actions in a changing world. The belief relations $\mathbf{B}(m)$ are reflexive and transitive and, therefore, Singh identifies belief with knowledge.

Definition 4.4.1.5 Semantic rules

1. $\mathcal{M}, t \models \mathsf{A}\varphi$ iff $(\forall \sigma \bullet \sigma \in \mathbf{S}_t \Rightarrow \mathcal{M}, \sigma, t \models \varphi)$
2. $\mathcal{M}, t \models \mathsf{E}\varphi$ iff $(\exists \sigma \bullet \sigma \in \mathbf{S}_t \Rightarrow \mathcal{M}, \sigma, t \models \varphi)$
3. $\mathcal{M}, t \models \mathsf{R}\varphi$ iff $\mathcal{M}, \sigma, t \models \varphi$, where $\sigma = \mathbf{R}(T)$
4. $\mathcal{M}, t \models (BEL\ m\ \varphi)$ iff $(\forall t' \bullet (t, t') \in \mathbf{B}(a) \Rightarrow \mathcal{M}, t' \models \varphi)$
5. $\mathcal{M}, t \models (\bigvee x : \varphi)$ iff $(\exists a \bullet a \in \mathcal{B}$ and $\mathcal{M}, t \models \varphi|^x_a)$, where $\varphi \in \mathcal{L}$
6. $\mathcal{M}, \sigma, t \models m[a]\varphi$ iff $(\exists t' \in \sigma \bullet [\sigma; t, t'] \in [\![a]\!]^m) \Rightarrow (\exists t' \in \sigma \bullet [\sigma; t, t'] \in [\![a]\!]$ and $(\exists t'' \bullet t \prec t'' \preceq t'$ and $\mathcal{M}, \sigma, t'' \models \varphi))$
7. $\mathcal{M}, \sigma, t \models m\langle a\rangle\varphi$ iff $(\exists t' \in \sigma \bullet [\sigma; t, t'] \in [\![a]\!]^m$ and $(\exists t'' \bullet t \prec t'' \preceq t'$ and $\mathcal{M}, \sigma, t'' \models \varphi))$
8. $\mathcal{M}, \sigma, t \models m \langle\!|a|\!\rangle \varphi$ iff $(\exists t' \in \sigma \bullet [\sigma; t, t'] \in [\![a]\!]^m$ and $(\exists t'' \bullet t \prec t'' \preceq t'$ and $(\forall t''' \bullet t \prec t''' \preceq t''$ implies that $\mathcal{M}, \sigma, t''' \models \varphi)))$
9. $\mathcal{M}, \sigma, t \models \varphi$ iff $\mathcal{M}, t \models \varphi$, where $\varphi \in \mathcal{L}_s$

In words: the operators A, E, and R are used to denote that a formula holds, respectively, in *all* scenarios at the given moment, in *some* scenario at the given moment, and in the *real* scenario at the given moment. The meaning of $(\bigvee x : \varphi)$ is that there exists an action under which φ becomes true. In this way, the notion of choice is expressed formally: an agent may be able to do several actions, but would, in fact, choose to do one. The action symbol x typically occurs in φ and is replaced by the specific action which makes φ true.

The definition of the modalities for actions differs from the way they are defined in Chapter 2. In the model of Singh, moments are defined independently of specific actions. Time is not constrained to be discrete. Singh does not take as an assumption that all actions are of equal duration and synchronized to begin and end together.

For an action symbol a, an agent symbol m, and a formula φ, $m[a]\varphi$ holds on a given scenario σ and a moment t on it, iff, if m performs a on σ starting at t, then φ holds at some moment while a is being performed. The formula $m\langle a\rangle\varphi$ holds on a given scenario and a moment t on it, iff, m performs a on σ starting at t and φ holds at some moment while a is being performed. The formula $m \langle\!|a|\!\rangle \varphi$ holds on a scenario σ and a moment t on it if m performs action a starting at t and φ holds in some initial subperiod of the period over which a is done.

Constraint 4.4.1.6 Constraints on the model

1. $([\sigma; t_0, t_2], [\sigma; t_1, t_3] \in [\![a]\!]^m) \Rightarrow (t_0 \preceq t_1 \prec t_2 \Rightarrow t_2 = t_3)$
2. $[\sigma; t, t'] \in [\![a]\!]^m \Rightarrow (\forall t'' \bullet t \preceq t'' \prec t' \Rightarrow [\sigma; t'', t'] \in [\![a]\!]^m)$
3. $(\forall t \bullet (\exists \sigma \in \mathbf{S}_t \bullet (\exists t' \in \sigma \bullet t \prec t')))$
4. $(\forall t \in \mathbf{T}, m \in \mathcal{A}(t), \sigma \in \mathbf{S}_t \bullet (\exists t' \in \sigma, a \in \mathcal{B} \bullet [\sigma; t, t'] \in [\![a]\!]^m))$
5. $(\forall \sigma \bullet (\forall t, t' \in \sigma \bullet t \prec t' \Rightarrow (\exists t'', a_1, \ldots, a_n \bullet t' \preceq t'' \wedge [\sigma; t, t''] \in [\![a_1, \ldots, a_n]\!])))$
6. $(\forall t, t' \bullet t' \in \mathbf{R}(t) \Rightarrow \mathbf{R}(t') \subseteq \mathbf{R}(t))$
7. $(\forall t, t', t_1 \in \mathbf{T}, \sigma, \tau \in \mathbf{S}_t \bullet (t \prec t' \prec t_1 \wedge [\tau; t, t'] \in \sigma) \Rightarrow ([\tau; t, t_1] \in [\![a]\!]^m \Rightarrow (\exists t_0 \in \sigma \bullet [\sigma; t, t_0] \in [\![a]\!]^m)))$

In the first constraint it is guaranteed that each action has a unique termination point. Starting at any given moment, each action can be performed in at most one way on any given scenario. Constraint 2 allows us to talk of an agent's actions at any moment at which they are happening, not just where they begin. The sense of this constraint is that an agent can achieve a particular condition at the beginning of an action, as well as during performing the action. The third constraint states that there is always a scenario available along which the world may evolve. While time progresses, it is assumed by constraint 4 that every agent performs a particular action on any scenario, even if it is some kind of dummy action. This assumption ensures that time does not just pass by itself, and is needed to make the appropriate connections between time and action. Since in the formal model time is not discrete, it is needed to exclude models in which there are moments that would require infinitely long action sequences to reach. Constraint 5 guarantees, therefore, the reachability of moments. If a scenario is determined to be the real scenario, the appropriate suffix of that scenario should be the real scenario. This is regulated by constraint 6. The seventh constraint deals with the atomicity of basic actions. If an agent is performing an action over a part of a scenario, then he completes that action on that scenario. Intuitively, $[\tau; t, t_1] \in [\![a]\!]^m$ means that m is performing a from t to t_1 on scenario τ. Therefore, he must be performing a in any subperiod of that, including $[\tau; t, t']$, which is the same as $[\sigma; t, t']$. Thus, a must be completed on σ.

4.4.2 Strategies

Strategies are abstract descriptions of behavior and may be seen as deterministic programs. The set of strategies, $\mathcal{L}_y$, is recursively defined as follows.

Definition 4.4.2.1 Syntactic rules for strategies

1. $\mathbf{skip} \in \mathcal{L}_y$
2. If $\varphi \in \mathcal{L}$ then $\mathbf{do}(\varphi) \in \mathcal{L}_y$
3. If $Y_1, Y_2 \in \mathcal{L}_y$ then $Y_1; Y_2 \in \mathcal{L}_y$
4. If $\varphi \in \mathcal{L}, Y_1, Y_2 \in \mathcal{L}_y$ then $\mathbf{if}\ \varphi\ \mathbf{then}\ Y_1\ \mathbf{else}\ Y_2 \in \mathcal{L}_y$
5. If $\varphi \in \mathcal{L}, Y_1 \in \mathcal{L}_y$ then $\mathbf{while}\ \varphi\ \mathbf{do}\ Y_1 \in \mathcal{L}_y$

The strategy $\mathbf{do}(\psi)$ denotes an abstract action, namely, the action of achieving ψ. Thus the main difference between strategies and deterministic programs is that the former are composed of abstract descriptions for achieving a condition, while the latter are composed of a finite alphabet of basic action symbols. The models for interpretation are extended with a function $\mathbf{Y}$ that assigns a strategy to every agent at each moment.

Definition 4.4.2.2 Models for interpretation
$\mathcal{M} = \langle \mathbf{T}, \prec, [\![\,]\!], \mathbf{Y}, \mathbf{B}, \mathbf{R} \rangle$ is a model for interpretation of formulae of $\mathcal{L}$ if

1. $\mathcal{M}$ satisfies the requirements mentioned in definition 4.4.1.2, and
2. $\mathbf{Y} : \mathcal{A} \times \mathbf{T} \mapsto \mathcal{L}_y$ is a function that assigns a strategy to every agent at each moment.

In order to extend the semantics to deal with strategies, I will now define the interpretation of strategies. Informally, $[\![Y]\!]^m$ will denote the set of periods over which Y is successfully performed by agent m.

Definition 4.4.2.3 Interpretation of strategies

1. $[\sigma; t, t'] \in [\![\mathbf{skip}]\!]$ iff $t = t'$
2. $[\sigma; t, t'] \in [\![\mathbf{do}(\varphi)]\!]$ iff $\mathcal{M}, t' \models \varphi$ and $(\forall t'' \bullet t \preceq t'' \prec t' \Rightarrow \mathcal{M}, t'' \not\models \varphi)$
3. $[\sigma; t, t'] \in [\![Y_1; Y_2]\!]$ iff $(\exists t'' \bullet t \preceq t'' \preceq t'$ and $[\sigma; t, t''] \in [\![Y_1]\!]$ and $[\sigma; t'', t'] \in [\![Y_2]\!])$
4. $[\sigma; t, t'] \in [\![\mathbf{if}\ \varphi\ \mathbf{then}\ Y_1\ \mathbf{else}\ Y_2]\!]$ iff $(\mathcal{M}, t \models \varphi$ and $[\sigma; t, t'] \in [\![Y_1]\!])$ or $(\mathcal{M}, t \not\models \varphi$ and $[\sigma; t, t'] \in [\![Y_2]\!])$
5. $[\sigma; t, t'] \in [\![\mathbf{while}\ \varphi\ \mathbf{do}\ Y_1]\!]$ iff $(t = t'$ and $\mathcal{M}, t \not\models \varphi)$ or $(\exists t_0 \ldots t_n \bullet t = t_0$ and $t' = t_n$ and $(\forall l \bullet 0 \leq l < n \Rightarrow ([\sigma; t_l, t_{l+1}] \in [\![Y_1]\!]$ and $\mathcal{M}, t_l \models \varphi))$ and $\mathcal{M}, t_n \not\models \varphi)$

4.4.3 Intentions

The intentions of an agent m correspond to the possible future courses of events that are preferred by m. The preferred scenarios must in some sense correspond to the strategies of m. So, the intentions of an agent have something to do with his strategies. Singh presents the following general definition of intentions: *An agent intends all the necessary consequences of his performing his strategy.*

The formal language, $\mathcal{L}$ is extended to include intentions. Four operators are added: $\langle\rangle_i$, $\langle\!\langle\rangle\!\rangle$, $*$, and $INTEND$. Informally we can say that $m\langle Y\rangle_i\varphi$ means that there is a period on which strategy Y is performed by agent m and at the end of that period φ will hold. Further $m * Y$ means that at moment t agent m has strategy Y. The operator $\langle\!\langle Y\rangle\!\rangle\varphi$ will be used to guarantee that on all scenarios on which the current strategy Y is performed at a moment in the future proposition φ will hold. Finally, $(INTEND\ m\ \varphi)$ means that agent m has a strategy and that strategy helps to bring about the result that φ holds.

Definition 4.4.3.1 Syntactic rules for intentions

1. If $\varphi \in \mathcal{L}_p$, $m \in \mathcal{A}$, and $Y \in \mathcal{L}_y$ then $m\langle Y\rangle_i\varphi \in \mathcal{L}_p$
2. If $\varphi \in \mathcal{L}_p$, $m \in \mathcal{A}$, and $Y \in \mathcal{L}_y$ then $\langle\!\langle Y\rangle\!\rangle\varphi \in \mathcal{L}_p$
3. If $m \in \mathcal{A}$ and $Y \in \mathcal{L}_y$ then $m * Y \in \mathcal{L}$
4. If $\varphi \in \mathcal{L}_p$ and $m \in \mathcal{A}$ then $(INTEND\ m\ \varphi) \in \mathcal{L}$

Convention 4.4.3.2 Abbreviation $m[Y]_i\varphi \stackrel{\text{def}}{=} m\neg\langle Y\rangle_i\neg\varphi$

Definition 4.4.3.3 Semantic rules for intentions

1. $\mathcal{M}, \sigma, t \models m\langle Y\rangle_i\varphi$ iff $(\exists t' \bullet [\sigma; t, t'] \in [\![Y]\!] \wedge \mathcal{M}, \sigma, t' \models \varphi)$
2. $\mathcal{M}, t \models m\langle\!\langle Y\rangle\!\rangle\varphi$ iff $\mathcal{M}, t \models \mathsf{A}(m\langle Y\rangle_i\mathsf{true} \rightarrow \mathsf{F}\varphi)$
3. $\mathcal{M}, t \models m * Y$ iff $\mathbf{Y}(m, t) = Y$
4. $\mathcal{M}, t \models (INTEND\ m\ \varphi)$ iff $(\exists Y \bullet \mathcal{M}, t \models m * Y \wedge m\langle\!\langle Y\rangle\!\rangle\varphi)$

Singh defines a few additional constraints in order to prevent some shortcomings.

Constraint 4.4.3.4 Additional constraints

1. $m * Y \Rightarrow \mathsf{E}a\langle Y\rangle_i\mathsf{true}$

2. $\mathcal{M}, t \models (INTEND\ m\ \varphi) \Rightarrow (\exists t' \bullet (t,t') \in \mathbf{B}(m) \wedge \mathcal{M}, t' \models \mathsf{RF}\varphi)$
3. $\mathcal{M}, t \models (INTEND\ m\ \varphi) \Rightarrow (\exists t' \bullet (t,t') \in \mathbf{B}(m) \wedge \mathcal{M}, t' \models \neg\mathsf{RF}\varphi)$
4. $\mathcal{M}, t \models (INTEND\ m\ \varphi) \Rightarrow (\forall t' \bullet (t,t') \in \mathbf{B}(m) \wedge \mathcal{M}, t' \models \mathsf{EF}\varphi)$

The first constraint ensures that if an agent has a strategy then that strategy is executable on some scenario. So, strategies are satisfiable, and therefore, intentions are satisfiable. The second constraint ensures that an agent's intentions are consistent with his beliefs. An agent's intentions are not ruled out by his beliefs. The third constraint makes it possible that an agent intends a condition φ and still does not fully believe that φ will be realized: the agent considers at least one alternative moment from which his intention is not realized. So it is possible for arbitrary $\mathcal{M}$ to intend φ at some moment and not to believe that φ will become true in the future on the real scenario. Brand [3, p.149] said it thus: "A person can intend without thinking that he will succeed." The fourth constraint ensures that an agent with an intention φ believes that he may succeed at least on one scenario. It is irrational for an agent to intend a condition φ and not willing to believe that he may succeed on any scenario.

In order to formalize *commitment* Singh introduces the decomposition of strategies in a first part and a second part. According to Singh, it is rational for an agent to persist with his intentions as long as he is succeeding with his strategy. Singh proposes a complicated constraint. Informally this constraint says that if an agent m has a non-trivial strategy, then he has after execution of the first part of his strategy the second part as his actual strategy. However, persistence with intentions is not a guarantee for success. An agent must also have the ability (or, in Singh's terminology know-how) to perform an action in the right way.

4.4.4 Evaluation

Singh has defined intentions and beliefs separately, and therefore undesired believed side-effects do not have to be intended by the agent. However, Singh has chosen for the closure property for intentions (C2). In his opinion, one cannot distinguish between a condition φ and its logical consequences. In my opinion, an agent who intends to study in Groningen does not have to intend to study in an arbitrary city with about 170,000 inhabitants, although the former (logically) implies the latter. As other drawbacks, the theory of Singh also validates the necessitation rule for intentions (C1) and unrestricted weakening (C7). Formally, an agent can have inconsistent intentions (C10), but in the theory of Singh if an agent intends both φ and ψ, he intends implicitly achieving them in some temporal order. No assumptions about time are made in this formalization of intentions.

Singh says that intentions are causes of actions by agents although he gives a totally different definition of intentions. I think it is correct to state that intentions are causes of actions, but in the formal system as handled by Singh intentions have a different function. The key concept in the framework is the strategy and intentions are based on strategies. According to Bratman [6] intentions play the following three functional roles.

- Intentions normally pose problems for the agent; the agent needs to determine a way to achieve them.
- Secondly, intentions provide a "screen of admissibility" for adopting other intentions.
- Furthermore, agents "track" the success of their attempts to achieve their intentions.

This conception of intentions totally differs from the one given by Singh.

In the formal framework of Singh intentions are based on strategies. It is in my opinion counterintuitive that the formation of strategies (abstract actions or *plans* in Pollack's terminology [37]) is not initiated by intentions. The problem is that execution of a strategy does not even have to be performed by the agent himself. The same strategy may be executed by another agent. The only conditions for a person to have some intentions are having a strategy and knowing that on some scenario that strategy will be executed by some agent.

In [49], Singh drops the concept of strategies and adds a component for intentions to his model. Model component I assigns to each agent a set of scenarios that the agent is interpreted as having selected or preferred. Intentions are now "the conditions that inevitably hold on each of the selected scenarios." While Singh needed in [48] four operators to model intentions, in [49] only one operator is defined.

Definition 4.4.4.1 New semantic rule

$\mathcal{M}, t \models (INTEND\ m\ \varphi)$ iff $(\forall \sigma \bullet \sigma \in I(m,t) \Rightarrow \mathcal{M}, \sigma, t \models \mathsf{F}\varphi)$

In his renewed formalization, Singh defines further two constraints to ensure satisfiability and persistence of the agent. The first constraint is straightforward. The second constraint states that if an agent selects some scenarios, then at future moments on those scenarios, he selects from among the future components of those scenarios.

Constraint 4.4.4.2 New constraints

1. $I(m,t) \neq \emptyset$
2. $(\sigma \in I(m,t)$ and $[\sigma; t, t'] \in [\![a]\!]^m) \Rightarrow (\forall \tau \bullet \tau \in I(m,t') \Rightarrow (\exists \upsilon \bullet \upsilon \in I(m,t)$ and $\tau \subseteq \upsilon))$

According to Haddadi [23], the theory of Singh stays at the theoretical level and does not provide much insight into the design and implementation of agents. Singh has adopted the *designer's perspective* for his theory of rational agents. In [12] it is argued that "adopting an *agent's perspective*, where the theory is to be used by the agent when reasoning about the world is more adequate."

4.5 Huang, Masuch, and Pólos

In this subsection I will discuss a preference-based modal action logic for agents with bounded rationality (called ALX) proposed by a research group from the University of Amsterdam (CCSOM). Preferences provide the basis for rational action. Goals are defined in terms of preferences and accessibilities. Preferences are relations between propositions. Because of the restrictions imposed upon the preference order, undesired properties of goals are ruled out. Huang et al. do not provide a separate operator for intentions.

4.5.1 The Formal Framework

In the formal language, $\mathcal{L}$, formulae are constructed using the ordinary logical operators. Further, they have $\langle a \rangle \varphi$ to denote the one-place existential accessibility relation for action a. The formula $\langle a \rangle \varphi$ is well known from propositional dynamic logic. $\mathbf{P}$ denotes the two-place preference relation. The two-place operator $\circ$ is in this context used for *updates*, which are changes caused by an action.

Definition 4.5.1.1 Syntactic rules

If $\varphi, \psi \in \mathcal{L}$ and a is an action, then $(\langle a \rangle \varphi)$, $(\varphi \circ \psi)$, and $(\varphi \mathbf{P} \psi)$ are formulae of $\mathcal{L}$.

Definition 4.5.1.2 Models for interpretation
$\mathcal{M} = \langle W, cw, \succ, \{R^a\}_a, V\rangle$ is a model for interpretation of formulae of $\mathcal{L}$ if

1. W is a set of possible worlds (states),
2. $cw : W \times \wp(W) \mapsto \wp(W)$ is a closest world function,
3. $\succ \subseteq \wp(W) \times \wp(W)$ is a comparison relation for preferences,
4. $R^a \subseteq W \times W$ is an accessibility relation for each action a, and
5. $V : \Phi \mapsto \wp(W)$ is an assignment function for primitive propositions.

The function cw determines a set of *closest* states relative to a given state, such that the new state fulfill some specified conditions, but resemble the old state as much as possible in all other aspects. It seems to be inconsistent that the authors use in their model both single worlds and sets of worlds (which they call *situations*). However, the authors aimed at modeling agents with bounded rationality. Agents with bounded rationality are less well informed. They may have an incomplete description of their actual state, incomplete knowledge of the accessibility relations, and an incomplete preference order over situations. So, in defining the semantic rules the authors depart from situations (sets of worlds) instead of single worlds.

Constraint 4.5.1.3 Constraints on the model

1. $cw(w, X) \subseteq X$
2. $w \in X \Rightarrow cw(w, X) = \{w\}$
3. $cw(w, X) \cap Y \subseteq cw(w, X \cap Y)$
4. $(\emptyset \not\succ X), (X \not\succ \emptyset)$
5. $cw(w, X \cap \bar{Y}) \succ cw(w, Y \cap \bar{X})$ and $cw(w, Y \cap \bar{Z}) \succ cw(w, Z \cap \bar{Y})$ implies $cw(w, X \cap \bar{Z}) \succ cw(w, Z \cap \bar{X})$, where $\bar{Y} = W \setminus Y$

The first constraint ensures that the closest φ-worlds are indeed φ-worlds. The second constraint ensures that w is its own (and unique) closest φ-world if φ is true at w. In the third constraint it is said that if the closest φ-worlds satisfy a condition ψ, then that worlds are also closest φ-and-ψ-worlds. In the last constraints the requirements of *normality* and *transitivity* are imposed on the preference relation. Normality stipulates that no comparison between two sets of worlds would involve an empty set of worlds.

Definition 4.5.1.4 Semantic rules

1. $[\![\ ]\!]_{\mathcal{M}} : \mathcal{L} \to \wp(W)$ is the interpretation function which is defined as follows.
   (a) $[\![p]\!]_{\mathcal{M}} = V(p)$
   (b) $[\![\neg\varphi]\!]_{\mathcal{M}} = W \setminus [\![\varphi]\!]_{\mathcal{M}}$
   (c) $[\![\varphi \wedge \psi]\!]_{\mathcal{M}} = [\![\varphi]\!]_{\mathcal{M}} \cap [\![\psi]\!]_{\mathcal{M}}$
   (d) $[\![\langle a\rangle\varphi]\!]_{\mathcal{M}} = \{w \in W : (\exists w' \in W \bullet (w, w') \in R^a \wedge w' \in [\![\varphi]\!]_{\mathcal{M}})\}$
   (e) $[\![\varphi \circ \psi]\!]_{\mathcal{M}} = \{w \in W : (\exists w' \in W \bullet w' \in [\![\varphi]\!]_{\mathcal{M}} \wedge w \in cw(w', [\![\psi]\!]_{\mathcal{M}}))\}$
   (f) $[\![\varphi \mathbf{P} \psi]\!]_{\mathcal{M}} = \{w \in W : cw(w, [\![\varphi \wedge \neg\psi]\!]_{\mathcal{M}}) \succ cw(w, [\![\psi \wedge \neg\varphi]\!]_{\mathcal{M}})\}$

2. $\mathcal{M}, w \models \varphi$ iff $w \in [\varphi]_{\mathcal{M}}$

The interpretation of $\langle a \rangle \varphi$ yields the set of worlds from where the agent can access at least one φ-world via action a. The interpretation of $(\varphi \circ \psi)$ yields the set of worlds where ψ holds so that one could have got there from a closest φ-world. The formula $\varphi \mathbf{P} \psi$ is true in those worlds in which the agent prefers the closest φ-and-not-ψ-worlds to the closest ψ-and-not-φ-worlds.

4.5.2 Goals as derived from Preferences

In [27] Huang et al. assume that the preference order of an agent, the corresponding set of preference statements, and the range of action alternatives $(a_1, \ldots, a_n)$ are finite.

Definition 4.5.2.1 Auxiliary definitions

1. $Acc(\varphi)$ iff $\langle a_1 \rangle \varphi \vee \cdots \vee \langle a_n \rangle \varphi$
2. $\mathbf{GO}\varphi$ iff $\varphi \mathbf{P} \neg \varphi$
3. $\mathbf{PO}\varphi$ iff $(\exists \psi \bullet (\varphi \mathbf{P} \psi) \vee (\psi \mathbf{P} \varphi))$

$Acc(\varphi)$ stands for the fact that situation φ is accessible via an action. If an agent prefers a situation φ to its negation, then we call φ a 'good' situation, $\mathbf{GO}\varphi$. Finally, $\mathbf{PO}\varphi$ stands for an element in the agent's preference order, which is possibly partial and irreflexive.

In ALX, goals are not a primitive notion as in the formalisms of Cohen and Levesque [9], and Rao and Georgeff[41]. Goals are derived from preferences. Huang et al. present us with four goal definitions: good goals, satisficing goals, maximal goals, and optimal goals.

Definition 4.5.2.2 Goals definitions

1. $GOAL^g(\varphi) \Leftrightarrow \mathbf{GO}\varphi$
2. $GOAL^s(\varphi) \Leftrightarrow \langle satisficing\_search \rangle \varphi \wedge \mathbf{PO}\varphi$
3. $GOAL^m(\varphi) \Leftrightarrow \mathbf{PO}\varphi \wedge (\forall \psi \bullet \psi \mathbf{P} \varphi \rightarrow \neg Acc(\varphi))$
4. $GOAL^o(\varphi) \Leftrightarrow GOAL^m(\varphi) \wedge (\forall \psi \bullet GOAL^m(\psi) \rightarrow (\varphi \leftrightarrow \psi))$

A *good goal* is a situation that is preferred to its negation. A *satisficing goal* is a satisfactory state that is made accessible via a search. The action $satisficing\_search$ is used to refer to an abstract action that consists of a satisficing search. So, this notation functions as a black box. A *maximal goal* is an accessible situation to which no other accessible situation is preferred, An *optimal goal* is a unique maximal goal. Note that in that case the preference order need be total.

Lemma 4.5.2.3 Avoidance of undesired properties

1. Goals are not closed under logical implication. Formally: $\models (\varphi \rightarrow \psi)$ does not imply $\models (GOAL(\varphi) \rightarrow GOAL(\psi))$
2. Goals do not satisfy the necessitation rule. Formally: $\models \varphi$ does not imply $\models GOAL(\varphi)$

4.5.3 Evaluation

No one of the criteria is validated by the formalism of Huang et al. But the theory lacks expressibility. The authors admit that ALX has several limitations. The action logic is restricted to a propositional description language. Further, the authors do not make use of a belief-operator. So, they cannot model the difference between objective and individual knowledge. ALX lacks an explicit notion of time and after all, the authors have only formalized single-agent acting. In [28] Huang presents an adapted formalism, ALX3. In ALX3 the logic is extended to overcome several shortcomings of ALX.

4.6 Van Linder, Van der Hoek, and Meyer

The last approach I examine in this chapter comes from a research group from Utrecht, the Netherlands. They formalize notions like preferences, goals, and commitments bottom up [33]. The notion of preferences forms the foundation for their formalism. Preferences consist of an implicit and an explicit part. Explicit preferences are just sentences of the language $\mathcal{L}$ that are assigned to each agent in a given world. Using the explicit preferences of the agent to act as a kind of filter on its implicit preferences allows one to avoid the side-effect and the transference problem. Goals are defined in terms of preferences. At the practition level, Van Linder et al. distinguish between a static aspect of commitment and a dynamic aspect. The former formalizes the commitments that agents have made. The latter formalizes the act of making commitments. For interpreting knowledge and implicit preferences the authors use a modal semantics.

4.6.1 The Formal Framework

In the language $K(\varphi)$ denotes the fact that the agent knows φ. In formalizing actions the authors use a slightly adapted variation of dynamic logic. Furthermore, the authors distinguish between *ability* and *opportunity* to do a particular action. The abilities of an agent comprise mental and physical powers, moral capacities, and human and physical possibility. The opportunity for an agent can be considered as circumstantial possibility. An example that may illustrate the difference between ability and opportunity is that of a lion in a zoo: although the lion will never have the opportunity to eat a zebra, it certainly has the ability to do so. The formula $\mathbf{A}(a)$ denotes that the agent has the ability to do a. The formula $\langle do(a)\rangle\varphi$ represents the fact that the agent has the opportunity to do a and that doing a leads to φ. $\mathbf{P}(\varphi)$ means that the agent prefers φ to be true. The formula $\Diamond\varphi$ has the intuitive interpretation that φ is implementable for the agent, i.e., there is some way open to the agent to bring about φ.

Definition 4.6.1.1 Syntactic rules
If $\varphi \in \mathcal{L}$ and a denotes some action, then $K(\varphi)$, $\mathbf{A}(a)$, $\langle do(a)\rangle\varphi$, $\mathbf{P}(\varphi)$, and $\Diamond\varphi \in \mathcal{L}$.

Definition 4.6.1.2 Models for interpretation
$\mathcal{M} = \langle W, \Pi, \mathcal{E}, r_0, c_0, P, Ep\rangle$ is a model for interpretation of formulae of $\mathcal{L}$ if

1. W is a set of possible worlds,
2. $\Pi : \Phi \times W \mapsto \{0,1\}$ assigns a truth value to propositional symbols in states,
3. $\mathcal{E} \subseteq W \times W$ is an epistemic accessibility relation,

4. $r_0 : \mathcal{B} \mapsto (W \times W)$ is a function that yields the state transition caused by a basic action,

5. $c_0 : \mathcal{B} \times W \mapsto \{0,1\}$ is a function that indicates whether the agent is capable of performing a particular basic action in a given world,

6. $P \subseteq W \times W$ is a preferential accessibility relation, and

7. $Ep : W \mapsto \wp(\mathcal{L})$ is a function that yields the explicit preferences of the agent.

The epistemic accessibility relation $\mathcal{E}$ is an equivalence relation , while the preferential accessibility relation P is serial. The functions r_0 and c_0 are extended to include composed actions using the constructs of dynamic logic and are referred to by r and c. These extended functions also account for special actions that transform models.

Definition 4.6.1.3 Semantic rules

1. $\mathcal{M}, w \models K(\varphi)$ iff $(\forall w' \in W \bullet (w, w') \in \mathcal{E} \Rightarrow \mathcal{M}, w' \models \varphi)$

2. $\mathcal{M}, w \models \langle do(\alpha) \rangle \varphi$ iff $(\exists \mathcal{M}', w' \bullet (\mathcal{M}', w') = r(\alpha)(\mathcal{M}, w) \wedge \mathcal{M}', w' \models \varphi)$

3. $\mathcal{M}, w \models \mathbf{A}(\alpha)$ iff $c(\alpha)(\mathcal{M}, w) = 1$

4. $\mathcal{M}, w \models \mathbf{P}(\varphi)$ iff $(\forall w' \in W \bullet ((w, w') \in P \Rightarrow \mathcal{M}, w' \models \varphi) \wedge \varphi \in Ep(w))$

5. $\mathcal{M}, w \models \Diamond \varphi$ iff $(\exists \alpha \bullet \mathcal{M}, w \models \langle do(\alpha) \rangle \varphi \wedge \mathbf{A}(\alpha))$

Constraint 4.6.1.4 Preferences persist and are known. For every action a:

1. $(\forall w, w' \in W \bullet w' = r(a)(w) \Rightarrow (\forall w'' \in W \bullet ((w', w'') \in P \Rightarrow (w, w'') \in P) \wedge Ep(w) \subseteq Ep(w')))$

2. $(\forall w, w' \in W \bullet (w, w') \in \mathcal{E} \Rightarrow (\forall w'' \in W \bullet ((w', w'') \in P \Rightarrow (w, w'') \in P) \wedge Ep(w) \subseteq Ep(w')))$

4.6.2 Goals and Commitments

The goals of an agent are determined by its unfulfilled preferences. Besides being unfulfilled, goals also need to be realistic, i.e., an unfulfilled preference is a goal for an agent if the agent knows that it is somehow possible to fulfill the preference.

Definition 4.6.2.1 Goal as an unfulfilled, realistic, known preference
$GOAL(\varphi) \stackrel{\text{def}}{=} KP(\varphi) \wedge K(\neg\varphi) \wedge K(\Diamond\varphi)$

Note that after performing an action that is correct with regard to a given goal, the goal ceases to be such. In formalizing commitments Van Linder et al. distinguish between a static and a dynamic part. The static component represents the commitments that agents have made. Models are accordingly extended with a new component: *Agenda*. This function returns for every state the actions the agent is committed to. The dynamic component consists in a language construct in a meta-language: *commit_to a*. The relation between the static and the dynamic component of commitments is as one would expect: when an agent (successfully) performs a *commit_to a* meta-action, this results in *committed a* (as part of the agenda) being true.

Table 4.1: Expressibility of the formalisms

	C&L	R&G	K&P	Singh	HM&P	LH&M
Multi agents				+		
First order logic	+	+				
Belief operator	+	+	+			
Knowledge operator	+			+	+	+
Goal operator	+	+			(+)	(+)
Intention operator	(+)	+	+	(+)		
Preference operator					+	+
Commitment operator						+
Time operator(s)	+	+		+		
Action operator(s)	+	+		+	+	+
Ability operator				+		+
Possibility operator	(+)	+	+		(+)	+

4.6.3 Evaluation

One of the main shortcomings in the formalism of Van Linder is, in my opinion, the fact that preferences are defined in a rather robust way. Preferences cannot be linked to each other. So, as a trivial consequence, undesired side-effects are avoided. I think, however, that some desired properties are also thrown away. For example, the way in which goals are linked to each other is not straightforward. Further, it is not clear how preferences and goals may change over time.

4.7 Conclusions

In this section I will provide three tables in order to summarize my investigations. In the first table I compare the expressibility of the various formalisms. The second table contains the classification of the temporal structure used for the formalisms. In the third table, I have collected the results of the test concerning the undesired properties discussed in Chapter 3

In Table 4.1 the expressibility of the discussed formalisms is reviewed. Parentheses in the table denote that the operator is defined in terms of other language components. The operator *Acc* in the formalism of Huang et al. resembles the possibility operator, usually denoted as $\Diamond$. $Acc(\varphi)$ stands for the fact that a situation φ is accessible via one of the action alternatives. In my opinion, *Acc* functions thus as a kind of possibility operator. In the formalisms of Cohen and Levesque, and Rao and Georgeff the possibility operator is used in the temporal sense: $\Diamond\varphi$ is used to denote that somehow or other φ will hold sometime in the future. In my opinion, this causes confusion. The possibility operator $\Diamond\varphi$ refers to a logical possibility and does not necessarily imply that φ will eventually become true.

In Table 4.2 I bring together the properties of the temporal logics underlying the various formalisms discussed above. I will use the classification criteria mentioned in Chapter 2. So, one can distinguish, respectively, between propositional and first order logics (T1), between endogeneous and exogeneous (T2), between branching and linear time (T3), between evaluation on points and evaluation on intervals (T4), between discrete and continuous models of time (T5), and finally between past and future tense (T6). Because Konolige and Pollack, Huang et al., and Van Linder et al. do not make explicit use of a temporal structure, I will

Table 4.2: Temporal structure of the formalisms

	C&L	R&G	Singh
T1	First order	First order	Propositional
T2	Endogeneous	Endogeneous	Endogeneous
T3	Linear time	Branching time	Branching time
T4	Points	Points	Points
T5	Discrete	Discrete	Continuous
T6	Past/Future	Past/Future	Past/Future

Table 4.3: Occurrence of undesired properties

	C&L	R&G	K&P	Singh	HM&P	LH&M
C1	No	Yes	No	Yes	No	No
C2	Yes	Yes	No	Yes	No	No
C3	Yes (1)	Yes	Yes (1)	No	No	No
C4	No	No	No	No (3)	-	No (3)
C5	No	No	No	No (3)	-	No (3)
C6	No	No	No	No (3)	-	No (3)
C7	Yes	Yes	No	Yes	No	No
C8	No	No	No	No (3)	-	No (3)
C9	No	Yes (2)	No	No (3)	-	No (3)
C10	No	No	No	Yes	No	No

omit them in this table.

In the last table of this chapter, Table 4.3, I bring together the results of the test: which undesired properties are validated by the above discussed formalisms? I refer the reader to look at Chapter 3 again and inspect the range of undesired properties, which I have used as criteria in order to compare the formalisms.

Although some authors use goals instead of intentions as the main motivational attitudes, I think that I can compare the motivational attitudes nevertheless. From the framework of Van Linder et al. I select as main motivational attitude the goal that the agent has himself committed to. From the framework of Huang et al. I select the *maximum achievement goal*, because it is the most urgent and realistic goal, and therefore similar to one's intention.

Three remarks are needed to clarify table 4.3. (1) In the formalisms of Cohen and Levesque, and Konolige and Pollack it is excluded that an agent at one and the same time intends both φ and an implication with φ as an antecedent. (2) This property is only validated if φ is an O-formula, i.e., φ contains no positive occurrences of A (*inevitable*) outside the scope of modal operators as BEL, $GOAL$, and $INTEND$. (3) In the formalisms of Singh and Van Linder et al., a knowledge operator instead of a belief operator is used. Since the accessibility relation for knowledge is an equivalence relation, the properties have to be interpreted differently.

Agent theories which are characterized by a high degree of expressibility seem to validate undesired properties more probably. Formalisms with a low degree of expressibility are sometimes too rigorous in constraining the various mental attitudes.

Chapter 5

From Theory to Practice

In computing science, *object-oriented* approaches have been used for the design, specification, and programming of conventional software systems. In object-oriented programming, a computational system may be seen as made up of modules that are able to communicate with one another. Objects encapsulate state information as a collection of data values and provide behaviors via well-defined interfaces for operators upon that information. However, "object-oriented methodologies are not directly applicable to agent systems - agents are usually significantly more complex than typical objects, both in their internal structure and in the behaviors they exhibit." [29, p.57]

In [46] Shoham presents an *agent-oriented programming* (AOP) framework. AOP can be viewed as a specialization of *object-oriented programming*. In this programming paradigm agents are programmed in terms of the mentalistic, intentional notions that agent theorists have developed to represent the properties of agents. In AOP, messages are categorized into types corresponding to types of speech acts, which will be discussed in the next chapter. In the same way that the intentional stance is used to describe humans, it might be useful to use the intentional stance to program machines.

In this chapter I will describe two proposals of *agent languages* , namely AGENT-0 [46] and PLACA [50]. An agent language is a system that allows one to program hardware or software computer systems in terms of the concepts developed by agent theorists [55, p.18]. Those concepts are discussed in the previous chapters. Further, I will evaluate the described languages and discuss the possibility of programming human behavior.

5.1 The AGENT-0 Language

Shoham is one of the pioneers working within the new computational framework, called *agent-oriented programming*. A computation may be seen as a community of agents informing, requesting from, offering to, accepting from, rejecting, competing with, and assisting one another. Shoham wants a complete AOP system to include three primary components:

- A restricted formal language with clear syntax and semantics for describing mental states (subsection 5.1.1);
- An interpreted programming language in which to define and program agents, with primitive commands such as REQUEST and INFORM (subsection 5.1.2);
- An 'agentifier', converting neutral devices into programmable agents.

The third component is "somewhat mysterious" according to Shoham. Agentification refers to bridging the gap between the low-level machine process and the intensional level of agent programs. An agentifier is a translator with a description of a machine in the process language as input, and an intensional program as output. Naturally, it may be assumed that the three components are not developed separately. However, Shoham's AOP system is criticized concerning this requirement [38].

5.1.1 Formalizing the Mental State

The logic contains three modalities: belief, commitment , and ability. In Shoham's view, the actions of an agent are determined by his *decisions*, or *choices*. "Some facts are true for natural reasons, and other facts are true because agents decided to make them so. Decisions are logically constrained, though not determined, by the agent's *beliefs*; these beliefs refer to the state of the world (in the past, present, or future), to the mental states of other agents, and to the *capabilities* of this and other agents." [46, p.60] Shoham considers, however, *obligation* (or commitment) as primitive, and treats decision simply as obligation to oneself.

Definition 5.1.1.1 Syntactic rules

If t denotes a time proposition, m and n denote agents, and φ a particular condition, then $(BEL\ m\ t\ \varphi)$, $(OBL\ m\ n\ t\ \varphi)$, $(DEC\ m\ t\ \varphi)$, $(CAN\ m\ t\ \varphi)$, and $(ABLE\ m\ \varphi)$ are element of the language to describe agents. Actions are represented by the corresponding fact holding.

Definition 5.1.1.2 Informal semantics

1. $(BEL\ m\ t\ \varphi)$ means: At time t agent m believes that φ
2. $(OBL\ m\ n\ t\ \varphi)$ means: At time t agent m is committed to agent n about φ
3. $(DEC\ m\ t\ \varphi) \stackrel{\text{def}}{=} (OBL\ m\ m\ t\ \varphi)$ means: At time t agent m has decided that φ be true
4. $(CAN\ m\ t\ \varphi)$ means: At time t agent m is capable of φ
5. $(ABLE\ m\ \varphi) \stackrel{\text{def}}{=} (CAN\ m\ time(\varphi)\ \varphi)$ means: Agent m is capable of φ at the appropriate time.

In the following constraints a number of properties is assumed about the modalities. In this way the modalities are restricted to resemble their common sense counterparts to some extent.

Constraint 5.1.1.3 Assumptions about the modalities

1. $(\forall m,t \bullet \{\varphi|(BEL\ m\ t\ \varphi)\}$ is consistent)
2. $(\forall m,t \bullet \{\varphi|(\exists n \bullet (OBL\ m\ n\ t\ \varphi))\}$ is consistent)
3. $(OBL\ m\ n\ t\ \varphi) \Rightarrow (BEL\ m\ t\ ((ABLE\ m\ \varphi) \wedge \varphi))$
4. $(OBL\ m\ n\ t\ \varphi) \Leftrightarrow (BEL\ m\ t\ (OBL\ m\ n\ t\ \varphi))$
5. $\neg(OBL\ m\ n\ t\ \varphi) \Leftrightarrow (BEL\ m\ t\ \neg(OBL\ m\ n\ t\ \varphi))$

In the first pair of constraints it is assumed that both the set of beliefs and the set of obligations are internally consistent. Further, it is assumed that agents commit only to what they believe themselves capable of, and only if they really mean it. In the last pair of constraints it is assumed that agents are aware of their obligations.

Remark 5.1.1.4 The reader may think of these assumptions as reasonable and straightforward, but to control the consistency of a set of formulae is not as simple as one would think so. Checking consistency for unconstrained theories is a notoriously hard problem, either intractable (in the propositional case) or undecidable (in the first-order case). And what to do, if one has to modify his beliefs and make the set of beliefs consistent afterwards?

5.1.2 Programming Language

In order to program agents, a programming language and an interpreter are needed. A program may be seen as a sequence of commitment rules, preceded by a definition of the agent's capabilities and initial beliefs, and the fixing of the "time-grain". A time grain is just an interval between two clock ticks. The basic loop of the interpreter works as follows. Each agent executes the following two steps at regular intervals.

- Read the current messages, and update your mental state, including your beliefs and commitments.
- Execute the commitments for the current time, possibly resulting in further belief change.

The syntax of the AGENT-0 language is defined in BNF notation in Table 5.1 on page 54. In accordance with standard conventions, * denotes repetition of zero or more times. The capabilities are assumed to be fixed.

Actions may be conditional or unconditional. The conditions in conditional actions refer to the mental state of the agent. When the time comes to execute the action, the mental state *at that time* will be examined to see whether the mental condition is satisfied. Further, actions may be private or communicative. AGENT-0 has three types of communicative actions: informing, requesting, and canceling a request. If an agent refrains from a particular action, then he will not commit himself to that action.

A commitment to an action is made if both a mental condition and a message condition are satisfied. The message condition refers to the current incoming messages. A message contains the sender's name, the type of the message, and the content of the message. A mental condition consists of the believed facts or the actions that the agent has committed himself to.

5.2 The PLACA Language

The *PLAnning Communicating Agents (PLACA)* language was originally intended to address one severe drawback to AGENT-0: the inability of agents to plan, and communicate requests for action via high-level goals. In [50], Thomas gives more attention to agent programming than to agent theory. For the latter she refers to her PhD thesis.

5.2.1 Formalizing the Mental State

An agent's mental state consists of the following components: beliefs about the world, capabilities to perform actions and to achieve states of affairs, plans to perform actions and

Table 5.1: Syntax of the AGENT-0 language

$< program >$	::=	$timegrain :=< time >$
		$CAPABILITIES := (< action >< mntlcond >)*$
		$INITIAL\ BELIEFS :=< fact > *$
		$COMMITMENT\ RULES :=< commitrule > *$
$< commitrule >$	::=	$(COMMIT < msgcond >< mntlcond > (< agent >< action >)*)$
$< msgcond >$	::=	$< msgconj > \mid$
		$(OR < msgconj > *)$
$< msgconj >$	::=	$< msgptrn > \mid$
		$(AND < msgptrn > *)$
$< msgptrn >$	::=	$(< agent > INFORM < fact >)\mid$
		$(< agent > REQUEST < action >)\mid$
		$(NOT < msgptrn >)$
$< mntlcond >$	::=	$< mntlconj > \mid$
		$(OR < mntlconj > *)$
$< mntlconj >$	::=	$< mntlptrn > \mid$
		$(AND < mntlptrn > *)$
$< mntlptrn >$	::=	$(BELIEF < fact >)\mid$
		$((COMMIT < agent >) < action >)\mid$
		$(NOT < mntlptrn >)$
$< action >$	::=	$(DO < time >< privateaction >)\mid$
		$(INFORM < time >< agent >< fact >)\mid$
		$(REQUEST < time >< agent >< action >)\mid$
		$(UNREQUEST < time >< agent >< action >)\mid$
		$(REFRAIN < action >)\mid$
		$(IF < mntlcond >< action >)$
$< fact >$	::=	$(< time > (< predicate >< arg > *))$
$< time >$	::=	$< integer > \mid$
		$< timeconst > \mid$
		$(+ < time >< time >)\mid$
		$(- < time >< time >)\mid$
		$(x < integer >< time >)\mid$
		now
$< timeconst >$	::=	$m \mid h \mid d \mid y$
$< agent >$	::=	$< string > \mid$
		$< variable >$
$< predicate >$	::=	$< string >$
$< arg >$	::=	$< string > \mid$
		$< variable >$
$< variable >$	::=	$? < string > \mid$
		$?! < string >$

to achieve states of affairs, and intentions to bring about states of affairs. When asked to accept a new task, an agent must perform two kinds of reasoning: deliberating about *whether* to take on the new task, and about *how* to accomplish it. The result of the first deliberation are intentions, the result of the latter are planning activities. Intentions are sentences that the agent has chosen to work towards making true, and that the agent is committed to. It is required that whenever an agent adopts a new intention, he must eventually create a plan for achieving the intended result, or else give up the intention. Further, it is required that an agent's beliefs and intentions should change only as necessary to (1) make the changes specified by rules that are fired, or (2) maintain consistency of the mental state overall.

5.2.2 Programming Language

At the most basic level, a PLACA agent's computation consist of the following steps:

- Collect messages received from other agents
- Update your mental state as specified in the program
- If sufficient time remains before the next tick of the clock, refine your plans (if necessary, a next time cycle may be used)
- Begin execution of the action to be performed next and return to the first step

Note that the planning is left to the agent who has adopted the intention.

Definition 5.2.2.1 PLACA program components

1. An initial mental state
   (a) A list of capabilities
   (b) A (consistent) list of initial beliefs
   (c) A (consistent) list of initial intentions
   (d) An empty set of initial plans
2. A list of mental-change rules

The syntax of the PLACA language is defined in BNF notation in Table 5.2 on page 56. As in the previous section, * denotes repetition of zero or more times. Further, + denotes repetition of one or more times. The capabilities are again assumed to be fixed.

An *atomic-sentence* is an atomic, unquantified sentence from the language $\mathcal{L}$ used to describe the agent's environment, or a sentence variable. Each sentence of $\mathcal{L}$ is dated. A *ground-sentence* is an atomic, variable-free sentence from language $\mathcal{L}$. Further, *action* and *agent* are terms from $\mathcal{L}$ to denote actions and agents.

5.2.3 Evaluation

After comparing the two languages, I collected the following differences between the two programming languages. 1) In AGENT-0, it is assumed that a single iteration through the loop of the interpreter last less than the time grain. In PLACA, a single loop may last longer than one cycle. 2) In AGENT-0, the obligations (or commitments) are the primary concept. In PLACA, everything revolves around mental changes. 3) In AGENT-0, both the capabilities of agents and the requests agents make of each other may refer only to primitive

Table 5.2: Syntax of the PLACA language

$< program >$	::=	$(< capabilities >, < initbeliefs >,$
		$< initintentions >, (< mntlchangerule > *))$
$< capabilities >$	::=	$(CAPABILITIES(< action >< sentence >)*)$
$< initbeliefs >$	::=	$(BELIEFS < ground\_sentence > *)$
$< initintentions >$	::=	$(INTENTIONS < ground\_sentence > *)$
$< mntlchangerule >$	::=	$(< msgcond >, < mntlcond >, < mntlchanges >, < msglist >)$
$< msgcond >$	::=	$< message > \mid$
		$(NOT < msgcond >) \mid$
		$(AND < msgcond >< msgcond > +) \mid$
		$(OR < msgcond >< msgcond > +) \mid ()$
$< message >$	::=	$(TO < agent >, FROM < agent >,$
		$< msgtype >, < sentence >)$
$< msgtype >$	::=	$INFORM \mid$
		$REQUEST \mid$
		$UNREQUEST$
$< sentence >$	::=	$< atomic\_sentence > \mid$
		$< mntlatom >$
$< mntlcond >$	::=	$< mntlatom > \mid$
		$(NOT < mntlatom >) \mid$
		$(AND < mntlcond >< mntlcond > +) \mid$
		$(OR < mntlcond >< mntlcond > +) \mid ()$
$< mntlatom >$	::=	$< external\_atom > \mid$
		$< internal\_atom >$
$< external\_atom >$	::=	$(BEL < sentence >) \mid$
		$(INTEND < sentence >)$
$< internal\_atom >$	::=	$(CAN\_DO < action >) \mid$
		$(CAN\_ACHIEVE < sentence >) \mid$
		$(PLAN\_DO < action >) \mid$
		$(PLAN\_NOT\_DO < action >) \mid$
		$(PLAN\_ACHIEVE < sentence >)$
$< mntlchanges >$	::=	$(< changeatom > +) \mid ()$
$< changeatom >$	::=	$(ADOPT < external\_atom >) \mid$
		$(DROP < external\_atom >)$
$< msglist >$	::=	$(< message > +) \mid ()$

actions. In PLACA, one agent can, for example, request of another that some state of affairs be achieved.

When I compare above formalizations of the mental state with the theories in Chapter 4, I see that Thomas's formalization bears the most resemblance to them. Thomas explicitly admits that her work has been strongly influenced by the work of Cohen and Levesque, especially their ideas on the role of commitment to one's intentions. Further, I see a point of agreement with the theory of Konolige and Pollack on the element of plan elaboration. Shoham's *DEC* modality does not resemble one of the motivational attitudes formalized in the previous chapter. Shoham's *OBL* (and hence the derived *DEC* modality) reflects absolutely no motivation of the agent, and merely describes the actions to which the agent is obligated. The commitment rules resemble the *Agenda*-function of [33].

While some of the authors discussed in the previous chapter use temporal modal operators like F, U, G, etc. (which are discussed in Chapter 2), Shoham and Thomas use explicit dates. It is remarkable that while in the theories of the previous chapter there was no much place for (cap)ability, both Shoham and Thomas define capabilities as element of the programming language. However, only Thomas uses capabilities explicitly as mental condition.

Rao criticizes Shoham and Thomas, for "they do not provide a formal proof theory or justify how the data structures capture the model-theoretic semantics of beliefs, commitments, and capabilities." [38, p.52] In the same article, Rao provides an operational and proof-theoretic semantics of a language AgentSpeak(L). In AgentSpeak(L), the current state of the agent, which is a model of itself, its environment, and other agents, can be viewed as as its current belief state; states which the agent wants to bring about based on its external or internal stimuli can be viewed as desires; and the adoptions of programs to satisfy such stimuli can be viewed as intentions. As in AGENT-0 and PLACA, the beliefs, desires, and intentions of the agent are not explicitly represented as modal formulas. However, while Shoham and Thomas assign the modalities as data structures to an agent, Rao describes the modalities to an agent from a designer's point of view.

5.3 Discussion

Intelligent agents have been investigated by many researchers from both a theoretical specification perspective and a practical design perspective. However, there still remains a large gap between theory and practice. According to Rao [38], the main reason for this has been the complexity of theorem-proving or model-checking in these expressive specification logics. Hence, the implemented systems have tended to use the mental attitudes as data structures, rather than as modal operators. Will it ever be possible that theory and practice are unified?

Kinny et al. [29] make a plea for adequate agent-oriented methodologies and modeling techniques for systems of agents. Agent-oriented methodologies demand another approach as object-oriented methodologies. In agent-oriented methodologies, "the focus is on the end-point, that is to be reached, rather than the types of behaviors that will lead to the end-point, which are the primary emphasis of object-oriented methodologies." [29, p.61] The application domain is analyzed in terms of what needs to be achieved, and in what context.

For specifying an agent system, Kinny et al. they distinguish between two levels of abstraction. "Firstly, from the *external viewpoint*, the system is decomposed into agents, modeled as complex objects characterized by their purpose, their responsibilities, the services they perform, the information they require and maintain, and their external interactions. Secondly, from the *internal viewpoint*, the elements required by a particular agent architecture must be modeled for each agent." [29, p.58]

The reason for focusing the attention to the design of agent-oriented systems is that they are being tested and installed in safety-critical applications, such as air-traffic management, real-time network management, and power-system management. Therefore, it is needed that those systems are verified and validated [42][39]. Wooldridge presents us with a framework for formally reasoning about implemented systems [54].

Totally different from the approaches described above is the approach employed by Wagner [52]. Wagner claims that his concept of *vivid agents* is able to narrow the gap between agent theory and practical systems. Wagner proposes a model of an agent which is both logical and operational. In his model, he combines a knowledge base with action and reaction rules, which yields an executable specification of agents. A vivid agent is a software-controlled entity whose state is represented by a knowledge base, and whose behavior is represented by means of action and reaction rules.

Looking back to several attempts of AOP, Burkhard provides us arguments for open systems under the special viewpoint of AOP [8]. In his opinion, there are many meaningful choices for the design of AOP languages coming from modeling human like behavior. We shall learn to live with ambiguities in the field of AOP, and use it as a chance to have flexible systems. One agent can be a single program on a single machine, and another agent could be another program on another machine. In Burkhard's view, the special application area of agent-oriented languages could be the support of the communication between the agents residing on different computers, the local control of an agent, the support of local actions as links to other private programs, and the control and information transfer between agents.

In the agent theories discussed in the previous chapter, motivational attitudes like goals, desires, and intentions have an important role. In AOP, therefore, mental states and mental attributes are thought to be essential. Burkhard, however, thinks that we need agent programs more guided perhaps by tasks and duties instead of intentions and desires. "Standard routines can be performed by the programs, while important decisions are due to the humans." [8, p.295] Burkhard does not believe in the replacement of human decisions by computers. Any application program must be programmed in order to follow the official rules, while real life decisions sometimes require the violation of rules and laws. Nobody would accept a program which may violate rules and laws e.g. in traffic domains.

5.4 Conclusions

In this chapter, I discussed some attempts to bridge the gap between theory and practice. The agent theories discussed in Chapter 4 all present logics of intention, but there is no indication of how these logics might be realized in a computer system. The AOP paradigm is based on systems claimed to have intentions but there is no statement or justification of which properties the intentions have [53]. The agent theories are often too expressive to implement. It is, for example, not clear how to implement a possible-worlds semantics . Knowledge systems are often used for implementing beliefs. The implementation of motivational attitudes seems to be not very easy. Maybe, it is impossible that implemented systems adopt intentions and plans of their own. I think the problem is that human motivational attitudes are often implicit and have an implicit order, while in implemented systems everything has to be made explicit as can be seen in the syntax for AGENT-0 and PLACA. In my opinion, it is for good reason that Shoham calls the process of agentification somewhat mysterious. So, I think that implementing systems for agents is possible as long as the motivational attitudes are limited and can be made explicit as, for example, with airline reservation systems [46].

Chapter 6

Communication

In this chapter I will investigate how communication can be coupled to the discussion of motivational attitudes in Multi-Agent Systems. In the previous chapter, some aspects of Multi-Agent Systems have already come to the fore. In AOP often attention is paid to message passing between agents (or computers) in a distributed computing system. An agent can request some action from another agent, and inform other agents. In this chapter, I will look further to the interaction between rational agents and how this *external viewpoint* is connected to the *internal viewpoint*. Why do I discuss communicative actions in a paper on motivational attitudes? I hope to make this clear. However, I will firstly describe how individual attitudes can be extended to group attitudes.

6.1 Group Attitudes

In Multi-Agent Systems one can distinguish between individual attitudes and group attitudes. The former are discussed in chapters 2 and 3. The latter are really a separate concept and not simply the sum of individual attitudes. In this section, I will successively discuss informational group attitudes and motivational group attitudes.

6.1.1 Common and Distributed Knowledge

In their handbook on knowledge, Fagin et al. add three modal operators to epistemic logic in order to express the notions of common and distributed knowledge [19, p.23f]. One of these is used to express the fact that every member of a group G has knowledge of a condition φ. A group G has *common knowledge* of a fact φ if simultaneously everyone knows φ, everyone knows that everyone knows φ, everyone knows that everyone knows that everyone knows φ, and so on. A group G has *distributed knowledge* of a fact φ if the knowledge of φ is distributed among its members, so that by pooling their knowledge together, the members of the group can deduce φ, even though it may be the case that no member of the group individually knows φ. In the sequel, $\mathcal{E}_m$ denotes the epistemic accessibility relation for agent m and $K_m(\varphi)$ denotes the fact that agent m knows φ.

Definition 6.1.1.1 Syntactic rules

If $G = \{1, \ldots, k\}$ is a nonempty set of agents and φ is a well-defined formula of $\mathcal{L}$, then the following formulae are admitted.

1. $K_G^e(\varphi)$ (everyone in G knows φ)

2. $K^e_G(\varphi)$ (φ is common knowledge among the agents in G)

3. $K^d_G(\varphi)$ (φ is distributed knowledge among the agents in G)

Definition 6.1.1.2 Semantic rules

1. $\mathcal{M}, w \models K^e_G(\varphi)$ iff $(\forall m \in G \bullet \mathcal{M}, w \models K_m(\varphi))$

2. $\mathcal{M}, w \models K^c_G(\varphi)$ iff $(\forall i > 0 \bullet \mathcal{M}, w \models (K^e_G)^i(\varphi))$

3. $\mathcal{M}, w \models K^d_G(\varphi)$ iff $(\forall w' \bullet (w, w') \in \bigcap_{m \in G} \mathcal{E}_m \Rightarrow \mathcal{M}, w' \models \varphi)$

Sometimes common (or mutual) knowledge is defined as *fixed points*.

Definition 6.1.1.3 Fixed point definition
$\mathcal{M}, w \models K^c_G(\varphi)$ iff $(\forall m \in G \bullet \mathcal{M}, w \models K_m(\varphi \wedge K^c_G(\varphi)))$

Remark 6.1.1.4 In the literature on Multi-Agent Systems the concept of *mutual belief* is used instead of common knowledge. It is possible to define concepts like mutual belief and distributed belief in a similar way as above.

6.1.2 Collective Intentions

In the previous subsection, I showed how common and distributed knowledge can be analyzed in terms of individual knowledge in a natural way. According to Searle [45], collective intentions cannot be analyzed in terms of individual intentions. Collective intentional behavior and collective intentions are primitive phenomena. Searle criticizes proposals of *we-intentions* that can be reduced to individual intentions and mutual belief. Examples of collective intentional behavior are two persons moving a heavy object, an orchestra playing a symphony, a football team playing football, or a group of construction workers building a house. The common view of collective activity is that of a group of agents having an objective that they wish to achieve collaboratively and, therefore, needs to be coordinated. Searle provides us the following explanation for defining collective intentions as a separate concept not reducible to individual intentions and mutual belief [45, p.406].

> The reason that we-intentions cannot be reduced to I-intentions, even I-intentions supplemented with beliefs and beliefs about mutual beliefs, can be stated quite generally. The notion of a we-intention, of collective intentionality, implies the notion of *cooperation*. But the mere presence of I-intentions to achieve a goal that happens to be believed to be the same goal as that of other members of a group does not entail the presence of an intention to cooperate to achieve that goal. One can have a goal in the knowledge that others also have the same goal, and one can have beliefs and even mutual beliefs about the goal that is shared by the members of a group, without there being necessarily any cooperation among the members or any intention to cooperate among the members.

To demonstrate that collective intentions are really a separate concept, you may have to think again of the members of an orchestra. They all have the goal of playing an instrument and making music in the knowledge that they all have the same goal. But if they lack a collective intention, i.e., if they are not making the same music and if they are not cooperating very well, they will not amuse the audience. In the same way, if a football team lacks a collective intention, then they will certainly not win the game. Of course, it is assumed

that each agent has a sense of other agents as "more than mere conscious agents, indeed as actual or potential members of a cooperative activity." [45, p.414]

There are in my opinion a few problems that have to do with the above characterization of collective intentions. Firstly, what is the relationship between the collective intention and the individual intentions that are all in one way or another directed at the same objective? Secondly, how are the individual intentions that form a means to the collective intention related to one another? It seems that from a philosophical point of view Searle is perfectly right. However, it is not in advance clear how collective intentions from this point of view may be formalized.

So far no serious attempts to formalize collective intentions as a separate concept are made. In my opinion, the reason is that such a formalization is not feasible in the area of distributed computing. For formalizing collective (or joint) intentions, Wooldridge and Jennings have introduced the concepts of *joint commitment* and *convention* . "When a group of agents are engaged in a cooperative activity, they have a joint commitment to the overall aim, as well as individual commitments to the specific tasks that they have been assigned. This joint commitment is parameterized by a social convention, which identifies the conditions under which the joint commitment can be dropped, and also describes how the agent should behave towards fellow team members." [56, p.43]

6.2 Interaction between Intelligent Agents

As described above, a Multi-Agent System may have distributed knowledge and collective intentions or, at least, joint commitments. The agents in a Multi-Agent System perform both their own tasks and group tasks. Therefore, it is needed that agents interact effectively with other agents, for instance, to update their knowledge [32]. It should be taken as a premise that much of communication depends upon an agent's ability to recognize one another's intentions and plans. In this section, I discuss first a way of communication between agents: linguistic communication. In order to deal with linguistic communication in a formal framework, philosophers of language have developed a *speech act* theory [43][51]. I will describe the various speech acts and how they are connected to other human actions. Next, I will show how speech acts can be related with motivational attitudes.

6.2.1 Speech Acts

According to Searle, speech acts are the basic or minimal units of linguistic communication. A speech act is, for example, a sentence like "I will go to Atlanta next week" or "The Olympic Games of this year take place in Atlanta." To each speech act, four distinct kinds of acts are connected: utterance acts (uttering of the words), propositional acts (referring and predicating), illocutionary acts (stating, questioning, commanding, promising, etc.), and perlocutionary acts (the effects of illocutionary acts) [43, p.24f].

A speech act per se is usually seen to have two parts: an *illocutionary force*, *illoc*, and a *proposition*, *p*. The illocutionary force characterizes the kind of the speech act and the proposition states what the speech act is about. According to Vanderveken [51, p.203], there are five primitive illocutionary forces (Table 6.1). All other illocutionary forces can be derived from the primitive forces.

Speech acts may be seen as basic actions. "A theory of language is part of a theory of action." [43, p.17] Speech acts can change the state of the world or at least the mental states of speakers and hearers. Like other human actions, speech acts can succeed or fail. "The *conditions of success* of an illocutionary act are the conditions that must be obtained in a

Table 6.1: Primitive illocutionary forces

Assertive	:Assertions as "The door is shut"
Commissive	:Commitments conveyed in "I will shut the door"
Directive	:Imperative sentences as "Shut the door!"
Declaritive	:Performative utterances as "I name this door the Golden Gate"
Expressive	:Exclamatory sentences as "I like this door"

possible context of utterance in order that the speaker succeed in performing that act in that context." [51, p.198] The *condition of satisfaction* of an illocutionary act in the context of utterance is the truth of its propositional content in the world of that context. Thus, while the conditions of success are connected with the illocutionary force, the condition of satisfaction is connected with the propositional content. So, there are two sets of semantic values: the success- and the truth-values. In the following subsection, I will provide with syntactic and semantic rules for the use of speech acts.

6.2.2 Speech Acts and Motivational Attitudes

Speech acts and motivational attitudes are related because both can be causes of actions. Speech acts can be utterances of motivational attitudes (assertions, exclamations). On the other side, motivational attitudes can be changed by speech acts (promises, commands). For example, if you assert φ then it may be assumed that you believe (or even know) φ, and if you promise to do some action α then it must be the case that you also intend to do α.

In the following definitions, I will provide a possible extension of the languages as discussed in Chapter 4. Singh has introduced the concept of communication in his language in [48]. Cohen and Levesque incorporate linguistic actions in their framework in [11]. In the sequel, I will freely make use of the various language constructs defined in Chapter 4. Recall that $Acc(\varphi)$ stands for the fact that situation φ is accessible via an action. $Acc_m(\varphi)$ is used to denote that situation φ is accessible for agent m. The set of language constructs is augmented with $ILLOC$, the set of illocutionary forces, and $MESS$, the set of messages. Each message has an illocutionary force and a propositional content, and is sent by a speaker and received by a hearer.

Definition 6.2.2.1 Language constructs and syntactic rule

1. $ILLOC = \{\text{assertive}, \text{commissive}, \text{declarative}, \text{directive}, \text{expressive}\}$
2. $MESS = \{(illoc, p) | illoc \in ILLOC \text{ and } p \in \Phi\}$
3. If $m, n \in \mathcal{A}$, $illoc \in ILLOC$, and $p \in \Phi$, then $\mathbf{comm}(m, n, (illoc, p)) \in \mathcal{B}$

Definition 6.2.2.2 Semantic rules for success

1. $\mathcal{M}, \sigma \models \langle \mathbf{comm}(m, n, (\text{assertive}, p)) \rangle \text{true}$ iff $\mathcal{M}, \sigma \models (BEL\ m\ p)$
2. $\mathcal{M}, \sigma \models \langle \mathbf{comm}(m, n, (\text{commissive}, p)) \rangle \text{true}$ iff $\mathcal{M}, \sigma \models (INTEND\ m\ p) \wedge (BEL\ m\ Acc_m(p))$
3. $\mathcal{M}, \sigma \models \langle \mathbf{comm}(m, n, (\text{declarative}, p)) \rangle \text{true}$ iff $\mathcal{M}, \sigma \models (INTEND\ m\ p) \wedge Acc_m(p)$

4. $\mathcal{M}, \sigma \models \langle \mathbf{comm}(m, n, (\text{directive}, p)) \rangle \text{true}$ iff $\mathcal{M}, \sigma \models (INTEND\ m\ (INTEND\ n\ p)) \wedge (BEL\ m\ Acc_n(p))$

5. $\mathcal{M}, \sigma \models \langle \mathbf{comm}(m, n, (\text{expressive}, p)) \rangle \text{true}$ iff $\mathcal{M}, \sigma \models \text{true}$

In the third semantic rule, proposition $Acc_m(p)$ is used to state that agent m is in the right social or conventional position to make the declarative succeed. In the case of a directive sentence it is naturally assumed that the speaker is in a position of authority over the hearer. In the last semantic rule, it is said that the utterance of expressive sentences does not require a specific state of the world. Instead, the expressive sentences require a specific state of the mind.

It is possible to define a series of constraints on the use of illocutionary acts. For example, it is not meaningful that an agent m orders an agent n to do an action, while n would do that action in the normal course of events of his own accord. Further, sometimes it is not meaningful that an agent asserts what is generally known. It would also be desirable to impose constraints on the use of expressive sentences. They are generally not allowed to contradict the state of the world or intended states of the world, because in that case they can cause confusion among the agents of the group.

Definition 6.2.2.3 Semantic rules for satisfaction

1. $\mathcal{M}, \sigma \models \mathit{succeeds}(\mathbf{comm}(m, n, (\text{assertive}, p)))$ iff $\mathcal{M}, \sigma \models p$

2. $\mathcal{M}, \sigma \models \mathit{succeeds}(\mathbf{comm}(m, n, (\text{commissive}, p)))$ iff $\mathcal{M}, \sigma \models (\exists \alpha \bullet \mathbf{A}_m(\alpha) \wedge \langle do(\alpha) \rangle p) \wedge (INTEND\ m\ p)$

3. $\mathcal{M}, \sigma \models \mathit{succeeds}(\mathbf{comm}(m, n, (\text{declarative}, p)))$ iff
   $\mathcal{M}, \sigma \models [\mathbf{comm}(m, n, (\text{declarative}, p))]\ p$

4. $\mathcal{M}, \sigma \models \mathit{succeeds}(\mathbf{comm}(m, n, (\text{directive}, p)))$ iff
   $\mathcal{M}, \sigma \models (\exists \alpha \bullet \mathbf{A}_n(\alpha) \wedge \langle do(\alpha) \rangle p \wedge [(\mathbf{comm}(m, n, (\text{directive}, p)))](INTEND\ n\ p))$

The assertion of a proposition p is satisfied if and only if it is true. A commissive illocutionary act is satisfied if and only if its propositional content is nontrivially achieved in a future state of the world. It requires that the speaker intends to achieve that proposition. A declarative illocutionary act is satisfied if and only if after every performance the appropriate proposition is achieved. A command or directive sentence is said to be satisfied if and only if its propositional content is nontrivially achieved by the hearer. Since in the performance of expressive illocutionary acts the speaker does not aim to achieve a state of the world, I have not given conditions of satisfaction.

Another way to look at interaction between intelligent agents is specifying interaction or cooperation *protocols*. "The idea behind the cooperation protocols is to provide a framework for designing context-dependent patterns of dialogue specific to the requirements of the domain, to relate messages to their context, enable the communicating agents to keep track of what has been communicated with respect to that context and how the dialogue should proceed." [23, p.107]

6.3 Conclusions

In this chapter, I discussed two aspects of Multi-Agent Systems against the background of mental attitudes. Firstly, I described the extension of individual mental attitudes to group

mental attitudes. In addition to individual knowledge there is also distributed knowledge and common knowledge. Besides individual intentions, there are also collective intentions. However, while distributed and common knowledge can be analyzed in terms of individual knowledge, collective intentions are not reducible to individual intentions. Therefore, it seems impossible to formalize collective intentions in such a way that it can be implemented in distributed systems. The best one can do is to formalize concepts like a *convention* [56] or a *shared plan* in order to express the fact of a *joint commitment* towards a condition.

Secondly, I described the way agents can communicate to one another in a Multi-Agent System. For that purpose, speech acts are used. On the one hand, speech acts are just basic actions. Each agent can perform an arbitrary speech act during an interval. On the other hand, speech acts resemble motivational attitudes. They can move someone to act. In any case, speech acts can change the mental states of agents. Speech acts can be used to update one's knowledge, to request some information or actions, to change (the priority order in) one's motivational attitudes, etc. I have provided some syntactic and semantic rules, because when modeling autonomous agents in a multi-agent world, it is inevitable to formalize the way in which agents communicate with each other.

Chapter 7

Summary and Conclusions

In this paper I have shed light on the process of formalizing motivational attitudes of autonomous intelligent agents in a multi-agent world. After an introduction in the first chapter, I arranged in Chapter 2 a few logics that are used as building blocks for the formalization of intelligent behavior. In Chapter 3, I investigated the properties of motivational attitudes. I gave priority to the concept of intentions because, in my opinion, they form most properly the link between an agent's mental state (knowledge and belief) and his behavior (actions). After having provided the philosophical background of motivational attitudes I looked at some proposals of an agent theory in Chapter 4. It should be clear that mathematical reasoning contrasts with the properties defined earlier. On the one hand, there are theories with a high degree of expressibility which, however, validate some undesired properties. On the other hand, there are theories which do not validate much undesired properties. However, they are too much constrained to be realistic. One main shortcoming of the discussed agent theories is the lack of relating intentions (or other motivational attitudes) to one another.

In Chapter 5, I discussed some attempts to apply the concepts of agents (with motivational attitudes) to programming and program design. It seems to be impossible to program agents having motivational attitudes. A designer can describe motivational attitudes to an agent, but he can not easily implement them. The problem is that human motivational attitudes are often implicit and have an implicit order, while in an implemented system all components have to be explicit. The execution of a computer program is guided by known rules. Human behavior seems on the one hand also guided by rules. On the other hand, human behavior seems to be unpredictable on principle. Anyway, further research will tell. In Chapter 6, I came to the conclusion that the formalization of speech acts is inevitable when modeling autonomous agents in a multi-agent world. Besides, speech acts can change the mental state of agents and, therefore, their motivational attitudes.

For an adequate theory of autonomous agents in a multi-agent world, the following components are needed: at least one informational and at least one motivational attitude, branching-time logics, possible-worlds, the notion of (cap)ability, a deontic operator, a notion like 'degree of strength' or 'priority' concerning mental attitudes, distinction between success and failure of actions, formalization of interaction between agents, and an action operator. In the enumeration given below, I will explain why in my opinion these components are needed for an adequate agent theory. Indirectly, I provide what may be seen as the conclusions of this paper.

- Why formalizing informational attitudes? Informational attitudes, like knowledge and belief, are essential for agents because they form the background for intelligent behav-

ior. Behavior may be seen as transformation of the state of the world. An agent is, however, not intelligent if he does not possess information about the actual state of the world, if he does not know what actions are available to him, if he does not know which states of the world are desirable, etc.

- Why formalizing motivational attitudes? Motivational attitudes are extensively discussed in Chapter 3. They form the necessary link between the knowledge of an agent and the actions that the agent consciously performs. So, if we are to explain intelligent behavior, we need an account of the motivations that caused the behavior.

- Why using branching-time logics? Singh provides us with the following argumentation [48, p.51]:

  > For our purposes, branching-time approaches yield a natural framework for describing the behaviour of multiagent systems. This is because multiagent systems are composed of intelligent agents who have limited control on the future of the world and exercise their choices independently of each other. Our models must incorporate the different choices available to agents explicitly, if we are to represent and reason about those choices and their optimality in our framework. Indeed, any formal framework that is sufficiently powerful for this purpose must involve at least some notion of branching-time, implicit or explicit.

  Thus, branching-time captures the notion of *choice*. Each path represents the options or choice of action available to the agent.

- Why using possible worlds? Each agent can have its own view of the environment and of other agent's mental states which may not coincide with the actual environment nor the actual mental states of these agents. These different views of the world, due to the uncertainty or chance inherent in the environment, can be effectively modeled within a possible-worlds framework. Thus, possible worlds may capture the notion of *chance*. Different belief-accessible worlds represent the agent's lack of knowledge. Different goal-accessible worlds represent the agent's desires that depend on the environment. Different intention-accessible worlds represent the commitments that are made relative to an unpredictable environment.

- Why formalizing ability? On the one hand, if an agent intends to do something but lacks the ability to do so, then he will surely not succeed with his intention. On the other hand, if an agent has the ability to do some action but lacks the motivation, then in general he will not perform that action. So, if we are to predict intelligent behavior, we need firstly an account of the motivations that cause the behavior and secondly the ability of the agent to perform the actions that constitute his potential behavior.

- Why using deontic logic? Motivational attitudes represent the internal reasons an agent has for his behavior. However, sometimes there are reasons external to the agent for his doing some action. You may have to think of a convention or an order to do some action. An agent has to drop his intentions if they contradict some of his obligations. In deontic logic, sentences like "It is obligatory that ..." or "It is permitted that ..." are studied. Deontic logic is, however, not part of the agent theories discussed in Chapter 4. I think, it is a shortcoming.

- Why formalizing a 'degree of strength'? In order to compare mental attitudes, it is sometimes useful to assign a probability to them. When I have two contradicting beliefs or intentions I have to decide between them. In normal modal logics the alternatives are equally valuable. However, if I want to go on holiday and apart from that I intend to go to Greece, to go to Spain, and to go to Czech, then it would be desirable to have a priority order among those intentions. Besides, some intentions will be dropped easier than others. In Chapter 3 it is said that intentions involve some *measure of commitment*. I think, therefore, that it is needed to express this measure in a theory about intentions although it will not be easy. In [16] Dongha proposed to assign to every intention a commitment level as a measure of anticipated and invested resources in order to get a priority order.

- Why distinguishing between success and failure? Some authors use a success theorem: If an agent intends to do some action and he has the ability to do that action, then he will succeed with his intention. However, even a child knows that there are several reasons for not succeeding with one's intentions, even though one has the ability. For computer systems you may have to think of a power failure. "Temporal logic does not distinguish between the conditions that the agent is trying and failing to achieve and the conditions that fail to hold for any other reason." [49, p.50] In my opinion, the success theorem must be weakened: If an agent intends to do some action and he believes that he has the ability to do that action, then at least he will try to perform that action. The attempt to perform some action may last some time.

  If an agent has failed to do some action a, even though he intended to do a, then nobody can give an explanation of the agent having done action b (which equals the failure of doing a) against the background of his intentions, because the failure of doing a has nothing to do with the original intention. In the meantime, however, the agent has not been idle. Only Rao and Georgeff [41] have paid attention to the distinction between success and failure. Unfortunately, they make not clear what exactly is meant by a basic action: the failure or the successful performance?

- Why formalizing speech acts? In a multi-agent world agents communicate with each other. They interact with each other using requests, assertions, orders, promises, etc. Speech acts can change the mental state of other agents. If an agent asserts something (for example, a teacher in a classroom) then it may be assumed that other agents update their beliefs. If an agent promises another agent to do some action, then it may be assumed that the former intends to do that action and that that action is in the interest of the latter agent. So, when someone is modeling autonomous agents in a multi-agent world, then he has to formalize the way the agents interact with each other.

- Why formalizing action? As stated above, actions may be seen as transformations of the state of the world. If an agent performs some action, then it may be assumed that the agent knows that he performs that action. Further, it may be assumed that after performing the action the agent will know that he has done that action and that he will bring his beliefs into line with the new state of the world. As another point, if an agent intends to achieve a condition, then it may be assumed that after achieving that condition, the agent will stop intending that condition. So, actions necessarily have to do with the state of the world and, therefore, with the state of the mind. For agents are perceiving the world around them and that will necessarily cause a regular update of the mental state.

In conclusion, I hope that I have demonstrated that there are still enough shortcomings in the formalization of the behavior of intelligent agents. So, for the design of intelligent agents further research is needed. In my opinion, at least the following topics need to be further developed.

Firstly, human behavior should be investigated more specific in order to develop agent theories that avoid undesired properties. For example, it will be needed that the motivational attitudes of a person at a moment are related to one another (in some priority order). Further, it will be needed to add a deontic modal operator to the theories. For agents often have to perform some action without inner motivation. It would also be desirable to distinguish successful performance of actions from failures.

Secondly, the gap between theory and practice should be narrowed by providing a proof-theoretic semantics of the languages for programming agents. Agent theorists should pay more intention to the verification and validation of implemented systems. The implementation and the underlying theory are too often developed independently.

Thirdly, in order to let agents communicate with each other (possibly humans), agent theorists should investigate the various ways of communication. A protocol for cooperation that is only suitable for simple requests fails in fully representing human communication. Every manner of interaction has its own formalization and its own semantics.

Index

Bibliography

[1] Michel Aubé and Alain Senteni. Emotions as commitments operators: A foundation for control structure in multi-agents systems. In Walter Van de Velde and John W. Perram, editors, *Agents Breaking Away: Proceedings of the 7th European Workshop on MAAMAW*, LNAI 1038, pages 13–25. Springer-Verlag, 1996.

[2] John Bell. Changing attitudes. In Michael J. Wooldridge and Nicholas R. Jennings, editors, *Intelligent Agents: ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, LNAI 890, pages 40–55. Springer-Verlag, 1995.

[3] Myles Brand. *Intending and Acting.* MIT Press, Cambridge, Massachusetts, 1984.

[4] Michael E. Bratman. Davidson's theory of intention. In Bruce Vermazen and Merrill B. Hintikka, editors, *Essays on Davidson: Actions and Events*, pages 13–26. Oxford University Press, New York, 1985.

[5] Michael E. Bratman. *Intentions, Plans, and Practical Reason.* Harvard University Press, Cambridge, Massachusetts, 1987.

[6] Michael E. Bratman. What is intention? In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 15–31. MIT Press, Cambridge, Massachusetts, 1990.

[7] Filip Buekens. *De mens en zijn intenties.* Pelckmans, Kapellen, Belgium, 1994.

[8] Hans-Dieter Burkhard. Agent-oriented programming for open systems. In Michael J. Wooldridge and Nicholas R. Jennings, editors, *Intelligent Agents: ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, LNAI 890, pages 291–306. Springer-Verlag, 1995.

[9] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[10] Philip R. Cohen and Hector J. Levesque. Persistence, intention, and commitment. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 33–69. MIT Press, Cambridge, Massachusetts, 1990.

[11] Philip R. Cohen and Hector J. Levesque. Rational interaction as the basis for communication. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 221–256. MIT Press, Cambridge, Massachusetts, 1990.

[12] Michael da Costa Móra, José Gabriel Lopes, and Helder Coelho. Modeling intentions with extended logic programming. In Jacques Wainer and Ariadne Carvalho, editors, *Advances in Artificial Intelligence*, LNAI 991, pages 69–78. Springer-Verlag, 1995.

[13] Donald Davidson. *Essays on Actions and Events.* Oxford University Press, New York, 1980.

[14] James P. Delgrande. A framework for logic of explicit belief. *Computational Intelligence*, 11:47–88, 1995.

[15] Daniel C. Dennett. *The Intentional Stance.* MIT Press, Cambridge, Massachusetts, 1987.

[16] Paul Dongha. Toward a formal model of commitment for resource bounded agents. In Michael J. Wooldridge and Nicholas R. Jennings, editors, *Intelligent Agents: ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, LNAI 890, pages 86–101. Springer-Verlag, 1995.

[17] Ho Ngoc Duc. Logical omniscience vs. logical ignorance on a dilemma of epistemic logic. In Carlos Pinto-Ferreira and Nuno J. Mamede, editors, *Progress in Artificial Intelligence*, LNAI 990, pages 237–248. Springer-Verlag, 1995.

[18] E.A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B*, pages 996–1072. Elsevier Science Publishers, 1990.

[19] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about Knowledge.* MIT Press, Cambridge, Massachusetts, 1995.

[20] Dov M. Gabbay, C.J.Hogger, and J.A. Robinson, editors. *Handbook of Logic in Artificial Intelligence and Logical Programming*, volume 1. Oxford University Press, New York, 1993.

[21] Graça Gaspar and Helder Coelho. Where do intentions come from?: A framework for goals and intentions adoption, derivation and evolution. In Carlos Pinto-Ferreira and Nuno J. Mamede, editors, *Progress in Artificial Intelligence*, LNAI 990, pages 115–127. Springer-Verlag, 1995.

[22] Michael P. Georgeff and Anand S. Rao. The semantics of intention maintenance for rational agents. In Chris S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 704–710. Morgan Kaufmann Publishers, San Mateo, California, 1995.

[23] Afsaneh Haddadi. *Communication and Cooperation in Agent Systems.* LNAI 1056. Springer-Verlag, 1995.

[24] David Harel. *First-Order Dynamic Logic.* LNCS 68. Springer-Verlag, 1979.

[25] Gilbert Harman. Willing and intending. In Richard E. Grandy and Richard Warner, editors, *Philosophical Grounds of Rationality*, chapter 14. Oxford University Press, New York, 1986.

[26] Koen Hindriks. On deciding what to do next. In *Proceedings of NAIC '96*, (forthcoming).

[27] Z. Huang, M. Masuch, and L. Pólos. ALX, an action logic for agents with bounded rationality. *Artificial Intelligence*, 82:75–127, 1996.

[28] Zhisheng Huang. *Logics for Agents with Bounded Rationality.* PhD thesis, Universiteit van Amsterdam, The Netherlands, 1994.

[29] David Kinny, Michael Georgeff, and Anand Rao. A methodology and modelling technique for systems of BDI agents. In Walter Van de Velde and John W. Perram, editors, *Agents Breaking Away: Proceedings of the 7th European Workshop on MAAMAW*, LNAI 1038, pages 56–71. Springer-Verlag, 1996.

[30] Kurt Konolige and Martha E. Pollack. A representationalist theory of intention. In Ruzena Bajcsy, editor, *Proceedings of the Thirteenth International Conference on Artificial Intelligence*, pages 390–395. Morgan Kaufmann Publishers, San Mateo, California, 1993.

[31] D. Kozen and J. Tiuryn. Logics of programs. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B*, pages 789–840. Elsevier Science Publishers, 1990.

[32] B. Van Linder, W. Van der Hoek, and J.-J.Ch. Meyer. Communicating rational agents. In Bernhard Nebel and Leonie Dreschler-Fischer, editors, *KI-94: Advances in Artificial Intelligence*, LNAI 861, pages 202–213. Springer-Verlag, 1994.

[33] B. Van Linder, W. Van der Hoek, and J.-J.Ch. Meyer. Formalising motivational attitudes of agents: on preferences, goals, and commitments. In M. Wooldridge, J.P. Müller, and M. Tambe, editors, *Intelligent Agents II: Agent Theories, Architectures, and Languages*, LNAI 1037, pages 17–32. Springer-Verlag, 1996.

[34] Bernardus Van Linder. *Modal Logics for Rational Agents.* PhD thesis, Universiteit Utrecht, The Netherlands, 1996.

[35] John McCarthy. Ascribing mental qualities to machines. In Vladimir Lifschitz, editor, *Formalizing Common Sense: Papers by John McCarthy*, pages 93–118. Ablex Publishing Corporation, Norwood, New Jersey, 1990.

[36] D.F. Pears. Intention and belief. In Bruce Vermazen and Merrill B. Hintikka, editors, *Essays on Davidson: Actions and Events*, pages 75–88. Oxford university Press, New York, 1985.

[37] Martha E. Pollack. Plans as complex mental attitudes. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 77–103. MIT Press, Cambridge, Massachusetts, 1990.

[38] Anand S. Rao. AgentSpeak(L): BDI agents speak out in a logical computable language. In Walter Van de Velde and John W. Perram, editors, *Agents Breaking Away: Proceedings of the 7th European Workshop on MAAMAW*, LNAI 1038, pages 42–55. Springer-Verlag, 1996.

[39] Anand S. Rao. Decision procedures for propositional linear-time belief-desire-intention logics. In M. Wooldridge, J.P. Müller, and M. Tambe, editors, *Intelligent Agents II: Agent Theories, Architectures, and Languages*, LNAI 1037, pages 33–48. Springer-Verlag, 1996.

[40] Anand S. Rao and Michael P. Georgeff. Asymmetry thesis and side-effect problems in linear-time and branching time intention logics. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 498–504. Morgan Kaufmann Publishers, San Mateo, California, 1991.

[41] Anand S. Rao and Michael P. Georgeff. Modeling rational agents within a BDI-architecture. In James Allen, Richard Fikes, and Erik Sandewall, editors, *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484. Morgan Kaufmann Publishers, San Mateo, California, 1991.

[42] Anand S. Rao and Michael P. Georgeff. A model-theoretic approach to the verification of situated reasoning systems. In Ruzena Bajcsy, editor, *Proceedings of the Thirtheenth International Joint Conference on Artificial Intelligence*, pages 318–324. Morgan Kaufmann Publishers, San Mateo, Calfornia, 1993.

[43] John R. Searle. *Speech Acts: An Esssay in the Philosophy of Language.* Cambridge University Press, 1969.

[44] John R. Searle. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge University Press, 1983.

[45] John R. Searle. Collective intentions and actions. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 401–415. MIT Press, Cambridge, Massachusetts, 1990.

[46] Yoav Shoham. Agent-oriented programming. *Artificial Intelligence*, 60:51–92, 1993.

[47] Munindar P. Singh. A critical examination of the Cohen-Levesque theory of intentions. In *Proceedings of the Tenth European Conference on Artificial Intelligence*, 1992.

[48] Munindar P. Singh. *Multiagent Systems: A Theoretical Framework for Intentions, Know-how, and Communication.* LNAI 799. Springer-Verlag, 1994.

[49] Munindar P. Singh. Semantical considerations on some primitives for agent specification. In M. Wooldridge, J.P. Müller, and M. Tambe, editors, *Intelligent Agents II: Agent Theories, Architectures, and Languages*, LNAI 1037, pages 49–64. Springer-Verlag, 1996.

[50] S. Rebecca Thomas. The PLACA agent programming language. In Michael J. Wooldridge and Nicholas R. Jennings, editors, *Intelligent Agents: ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, LNAI 890, pages 355–370. Springer-Verlag, 1995.

[51] Daniel Vanderveken. On the unification of speech act theory and formal semantics. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 195–220. MIT Press, Cambridge, Massachusetts, 1990.

[52] Gerd Wagner. A logical and operational model of scalable knowledge- and perception-based agents. In Walter Van de Velde and John W. Perram, editors, *Agents Breaking Away: Proceedings of the 7th European Workshop on MAAMAW*, LNAI 1038, pages 26–41. Springer-Verlag, 1996.

[53] Wayne Wobcke. Plans and the revision of intentions. In Chengqi Zhang and Dickson Lukose, editors, *Distributed Artificial Intelligence: Architecture and Modelling. Proceedings of the First Australian Workshop on DAI*, LNAI 1087, pages 100–114. Springer-Verlag, 1996.

[54] Michael Wooldridge. This is MyWorld: The logic of an agent-oriented DAI testbed. In Michael J. Wooldridge and Nicholas R. Jennings, editors, *Intelligent Agents: ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, LNAI 890, pages 160–178. Springer-Verlag, 1995.

[55] Michael Wooldridge and Nicholas R. Jennings. Agent theories, architectures, and languages: A survey. In Michael J. Wooldridge and Nicholas R. Jennings, editors, *Intelligent Agents: ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, LNAI 890, pages 1–39. Springer-Verlag, 1995.

[56] Michael Wooldridge and Nicholas R. Jennings. Towards a theory of cooperative problem solving. In John W. Perram and Jean-Pierre Müller, editors, *Distributed Software Agents and Applications: Proceedings of the 6th European Workshop on MAAMAW*, LNAI 1069, pages 40–53. Springer-Verlag, 1996.